

The international journal of science / 28 November 2019

outlook
Vaccines

nature

TWISTED TALE

Inflammation drives formation of
tau tangles in Alzheimer's disease

Reef revivals

Strategies to help coral
cope with a rapidly
warming world

Conversion process

Catalyst on nanotubes
offers efficient way to
turn CO₂ into methanol

Attractive prospect

How bacteria can
benefit from chasing
non-nutritious signals

Vol. 375, No. 7584
nature.com

Protect Italy's new funding agency

Lawmakers must ensure that the proposed ANR is independent and autonomous.

Italy is a rare example of a major world economy without a research funding agency that operates independently of a science or research ministry. For years, ministers and civil servants have had a say in what research gets funded. But that could be about to change.

Prime Minister Giuseppe Conte's government says it is committed to creating an independent National Research Agency (ANR) and has begun drafting its structure and functions. About time, too, say researchers who are tired of insufficient funds coming from opaque funding mechanisms. That said, they are worried that some politicians might wish to become too closely involved in the new agency's affairs (see *Nature* 575, 424–425; 2019).

It's a valid concern, and the fledgling agency will need to be protected. Paradoxically, it is politicians who must perform this crucial task, especially Italy's team of research ministers – led by political economist Lorenzo Fioramonti. It won't be easy, and Fioramonti and his team must steel themselves for the task ahead.

Conte pledged to establish the ANR in September, in his first speech as leader of a new coalition between the populist Five Star Movement and the centre-left Democratic Party. He added that financial support for research would also be increased.

The ANR, Conte said, would be modelled on science funding agencies in other European countries, which operate under the broad principle that politicians decide how much to allocate for research and have a say in strategic funding priorities. However, politicians do not decide which proposals are funded; nor are they involved in setting criteria for awards, or in evaluation. These tasks need to be performed independently, by subject experts chosen by the research community.

Under a proposal that has been presented to Italy's parliament as part of the 2020 budget, the ANR will receive €25 million (US\$28 million) for 2020, then €200 million for 2021 and €300 million annually from 2022. These are small sums by the standards of similar-sized economies, but it's a start. The ANR will coordinate research at universities and public research institutes. It will also fund “highly strategic” projects, and encourage participation in international research initiatives and cooperation with the private sector.

But the fine print – or lack of it – is causing concern. The current draft law says that the ANR's nine-member governing board will be nominated by university presidents, as well as representatives from the prime minister's office and government ministries. This is an unusually high level of

“Researchers and their representative organizations have been left out of the loop.”

involvement from political representatives, fuelling fears that the agency will come under the influence of politicians.

In line with international best practice, the ANR should also appoint a network of independent research advisers, drawn from various disciplines, to oversee the quality of funding calls and of funded applications. But the proposal presented to parliament has no such provision.

Many researchers are undecided on whether to see the glass as half empty or half full. They welcome the new support for research, but some would have preferred for Italy's existing funding bodies to be made part of the new agency. They are also unhappy that researchers and their representative organizations have been left out of the loop.

That said, researchers do have two things on their side. First, Fioramonti has said publicly that he is unhappy with the draft law and has promised to convince parliament to amend it. Second, a preliminary screening of the budget law by Italy's court of auditors has highlighted the question of independence and the auditors have asked for more clarity on how this will be achieved.

Fioramonti has said he is determined to do what he can to establish a healthy distance between the ANR and politicians, to consult the scientific community and to ensure that the ANR is governed to the highest possible standards of quality and probity.

Italy's researchers, and particularly the future generations whose careers the ANR will fund, need the minister and his team to follow through on these promises. Independence, transparency and trust are vital for an agency that could shape research in Italy for decades to come.

Egypt and the Egyptologists

Scientists everywhere are keen to share in the excitement of discoveries from Egypt's past.

It's been a busy few months for Egyptologists. Last week, the discovery of a cache of mummified animals – including the remains of lion cubs – dating back to ancient Egypt's 26th Dynasty (664–525 BC) was announced at the Saqqara necropolis, south of Cairo.

Last month, officials revealed that 30 sealed coffins and their mummified human contents had been found in the Assasif necropolis near Luxor. These are thought to be linked to the Amun priesthood, one of ancient Egypt's centres of power, which dates back to the tenth century BC. Further discoveries made in the country will be announced next month, according to Egypt's antiquities minister, Khaled El-Enany.

Researchers attending the annual congress of the International Association of Egyptologists in Giza earlier this month told *Nature* of their excitement about

Protect Italy's new funding agency

Lawmakers must ensure that the proposed ANR is independent and autonomous.

Italy is a rare example of a major world economy without a research funding agency that operates independently of a science or research ministry. For years, ministers and civil servants have had a say in what research gets funded. But that could be about to change.

Prime Minister Giuseppe Conte's government says it is committed to creating an independent National Research Agency (ANR) and has begun drafting its structure and functions. About time, too, say researchers who are tired of insufficient funds coming from opaque funding mechanisms. That said, they are worried that some politicians might wish to become too closely involved in the new agency's affairs (see *Nature* 575, 424–425; 2019).

It's a valid concern, and the fledgling agency will need to be protected. Paradoxically, it is politicians who must perform this crucial task, especially Italy's team of research ministers – led by political economist Lorenzo Fioramonti. It won't be easy, and Fioramonti and his team must steel themselves for the task ahead.

Conte pledged to establish the ANR in September, in his first speech as leader of a new coalition between the populist Five Star Movement and the centre-left Democratic Party. He added that financial support for research would also be increased.

The ANR, Conte said, would be modelled on science funding agencies in other European countries, which operate under the broad principle that politicians decide how much to allocate for research and have a say in strategic funding priorities. However, politicians do not decide which proposals are funded; nor are they involved in setting criteria for awards, or in evaluation. These tasks need to be performed independently, by subject experts chosen by the research community.

Under a proposal that has been presented to Italy's parliament as part of the 2020 budget, the ANR will receive €25 million (US\$28 million) for 2020, then €200 million for 2021 and €300 million annually from 2022. These are small sums by the standards of similar-sized economies, but it's a start. The ANR will coordinate research at universities and public research institutes. It will also fund “highly strategic” projects, and encourage participation in international research initiatives and cooperation with the private sector.

But the fine print – or lack of it – is causing concern. The current draft law says that the ANR's nine-member governing board will be nominated by university presidents, as well as representatives from the prime minister's office and government ministries. This is an unusually high level of

“Researchers and their representative organizations have been left out of the loop.”

involvement from political representatives, fuelling fears that the agency will come under the influence of politicians.

In line with international best practice, the ANR should also appoint a network of independent research advisers, drawn from various disciplines, to oversee the quality of funding calls and of funded applications. But the proposal presented to parliament has no such provision.

Many researchers are undecided on whether to see the glass as half empty or half full. They welcome the new support for research, but some would have preferred for Italy's existing funding bodies to be made part of the new agency. They are also unhappy that researchers and their representative organizations have been left out of the loop.

That said, researchers do have two things on their side. First, Fioramonti has said publicly that he is unhappy with the draft law and has promised to convince parliament to amend it. Second, a preliminary screening of the budget law by Italy's court of auditors has highlighted the question of independence and the auditors have asked for more clarity on how this will be achieved.

Fioramonti has said he is determined to do what he can to establish a healthy distance between the ANR and politicians, to consult the scientific community and to ensure that the ANR is governed to the highest possible standards of quality and probity.

Italy's researchers, and particularly the future generations whose careers the ANR will fund, need the minister and his team to follow through on these promises. Independence, transparency and trust are vital for an agency that could shape research in Italy for decades to come.

Egypt and the Egyptologists

Scientists everywhere are keen to share in the excitement of discoveries from Egypt's past.

It's been a busy few months for Egyptologists. Last week, the discovery of a cache of mummified animals – including the remains of lion cubs – dating back to ancient Egypt's 26th Dynasty (664–525 BC) was announced at the Saqqara necropolis, south of Cairo.

Last month, officials revealed that 30 sealed coffins and their mummified human contents had been found in the Assasif necropolis near Luxor. These are thought to be linked to the Amun priesthood, one of ancient Egypt's centres of power, which dates back to the tenth century BC. Further discoveries made in the country will be announced next month, according to Egypt's antiquities minister, Khaled El-Enany.

Researchers attending the annual congress of the International Association of Egyptologists in Giza earlier this month told *Nature* of their excitement about

the discoveries (see page 573). But some also expressed disappointment that Egypt's government will be restricting access – at least for now – to researchers at Egyptian institutions. There will be no open calls for research proposals of the type that museums and funding agencies typically publish to attract the best ideas and expertise.

The government has justifiable reasons for being careful about permitting further international involvement in its heritage. During colonial times, some of Egypt's most precious artefacts were taken, and many have wound up in Europe's leading museums.

The last time that coffins and mummies were discovered on a large scale was in 1891, at Bab el-Gasus ('the door of the priests'), not far from Luxor. Some of the surviving coffins from that find are now at the Rijksmuseum van Oudheden in Leiden in the Netherlands, and at the Vatican. Moreover, Zahi Hawass, Egypt's former antiquities minister, has long called for the return of the Rosetta Stone, which has been at the British Museum in London for more than 200 years.

But today's Egyptology bears little relation to the field's earlier era. Egypt hosts hundreds of teams of archaeologists from museums and universities around the world who are working in partnership with Egypt's universities and government. At last month's congress for Egyptologists, both Hawass and El-Enany were among the main speakers.

There are also many models for research collaboration. Egypt could, for example, issue calls for proposals in which international researchers are invited to join Egypt-led research consortia as co-investigators.

Every nation is the custodian of its heritage – a right that must never again be taken away. But at the same time, Egypt's rich history, which encompasses many civilizations, is also an example of how science and scholarship flourish when there are few barriers to talent. That is why, when Egypt feels the time is right, its government should consider inviting more of the world's researchers to work with its own, allowing them to contribute to the latest finds from the country's fascinating past.

“
Science and
scholarship
flourish
when there
are few
barriers
to talent.”

Troubling trends

Attacks on scholars are on the rise at the same time as universities in several countries find themselves at the centre of student protests.

“**S**owing corruption on Earth”. That was one of the charges levelled at Iranian conservation biologists who were arrested in January 2018 and charged with spying. They were arrested for using camera traps to study endangered wildlife, especially the Asiatic cheetah (*Acinonyx jubatus venaticus*). There are fewer than 100 of the animals left in the world and most are believed to be in Iran.

All the nine researchers charged – Niloufar Bayani,

Taher Ghadirian, Amirhossein Khaleghi Hamidi, Houman Jowkar, Sepideh Kashani, Abdolreza Kouhpayeh, Sam Rajabi, Morad Tahbaz and Kavous Seyed Emami – were associated with the Persian Wildlife Heritage Foundation, a well-known Tehran-based wildlife conservation charity that had strong links to international conservation organizations and to the UN Environment Programme.

Emami died in unexplained circumstances in prison shortly after his arrest. The other eight were sentenced last week to between 6 and 10 years in prison, but are strongly protesting their innocence. The trial was held in secret, despite an international outcry from leading conservation charities and pleas from the United Nations for a fair and transparent process.

This tragic verdict came too late for inclusion in *Free to Think 2019*, an annual report from Scholars at Risk, an international organization that highlights human-rights violations against academic researchers and students. Now in its fifth year, the report records the experiences of scholars who have been subjected to violent or fatal attacks, wrongful prosecution or imprisonment, or who have been sacked or expelled from their institution without undergoing due process.

It isn't only in Iran that the law is being misused in such a way. The Scholars at Risk report highlights cases of rights violations in 56 countries. This year's tally of 324 recorded cases between 1 September 2018 and 31 August 2019 is higher than last year's 294, although the report points out that the examples are just a snapshot of a larger picture.

And there is another emerging phenomenon that has come too late to be highlighted in this year's study, but is likely to appear in the next. This is the scenes of campus unrest in such diverse locations as Chile, Hong Kong, Iran, Iraq and Lebanon.

Night after night, thousands of young people, as well as their teachers and lecturers, are taking to the streets or – in Hong Kong's case – have been protesting inside university campuses. Campuses are often places for dissent, but what is happening now is on a scale rarely seen in recent times.

These protests are often in response to a lack of jobs, rising prices, falling living standards, environmental concerns, or concerns about weak, unrepresentative or corrupt political leadership.

The response from university leadership depends on the context. In Iran's case, speaking out is not an option. In Lebanon, where there is much more academic freedom, students and lecturers have organized informal teach-ins and university presidents are calling on political leaders to heed their students' demands.

As 2019 gives way to 2020, it is unlikely that campus unrest will abate. There will be pressure from governments on university management not to allow premises to be used for demonstrations. And there will be pressure from the academic community and students not to give in to these, and more draconian, demands.

Researchers and students should not need to live in fear in the pursuit of their science. As this year's Scholars at Risk report demonstrates, that more are having to do so is a troubling trend.

World view

Africa should set its own health-research agenda



By Francisca Mutapi

Local experts – not rich donors – must design and control studies, says Francisca Mutapi.

In a controversy reported last month, critics accused a UK institute of using African people's DNA inappropriately, without sharing the benefits with partner institutions in Africa. But the biggest failing I see in transcontinental partnerships goes deeper. It involves inequity in the control of funding, research agendas, outputs, training and infrastructure. At a meeting I led in Accra, Ghana, this year, funders, policymakers and researchers agreed: 'safari science' is ineffective. Inequitable partnerships that task African scientists as data gatherers for Western research agendas are unlikely to make a difference to the African health problems that really matter.

I've seen this play out for decades. For more than 20 years, I have led a programme in Zimbabwe on human schistosomiasis. For most of that time, international donors concentrated on treating schoolchildren. Our team's persistence led to the extension of treatment and monitoring to pre-school children, a policy now endorsed by the World Health Organization (WHO). In 2017, I began co-leading a new UK-funded partnership, Tackling Infections to Benefit Africa (or TIBA, the Swahili word for 'curing infection').

TIBA brings together world-class researchers from Botswana, Ghana, Kenya, Rwanda, South Africa, Sudan, Tanzania, Uganda and Zimbabwe, plus colleagues at the University of Edinburgh, UK. One project studies the Chikungunya virus, a global public-health problem that is largely ignored in Africa – but that our researchers found was linked to almost 30% of fever cases at one Kenyan hospital. Another examines people who are asymptomatic carriers of sleeping sickness, an under-researched problem that could thwart efforts to eliminate the disease in Uganda.

Four principles are crucial for TIBA. First, our research activities are led from Africa, chosen to reflect local priorities and not dictated by outside agencies. One example is our work on the autoimmune disease systemic lupus erythematosus, which is more common and more severe in people of African descent. The international diagnostic criteria were derived from non-African people presenting mainly with inflamed joints and mucosal membranes. A study of affected Zimbabweans identified a variant of the disease in Africans that is characterized mainly by rashes and skin lesions (Sibanda, E. N. *et al. BMJ Glob. Health* 3, e000697; 2018).

Second, we are shifting the centre of gravity for African health-research decisions to Africa (where it belongs). The bulk of our work – and 80% of our spending – takes place in Africa. I have seen too many projects in which most research funds go to laboratories in the global north or to the salaries of expatriate researchers.


'Safari science' is ineffective.'

Francisca Mutapi is deputy director of the National Institute for Health Research's Global Health Research Unit Tackling Infections to Benefit Africa at the University of Edinburgh, UK.
e-mail: f.mutapi@ed.ac.uk

Third, we strive to be equitable. African experts form most of our directorate, steering committee and external advisory group. African-based researchers are the first authors on 9 of the 14 papers we have published in the past 2 years.

Fourth, we aim for inclusivity. Each partner engages stakeholders – from affected communities to national health ministries – at the outset of each project. All of our partners have access to our training and capacity-building activities. We explicitly ask how outputs from a project will guide local decision-making and benefit local populations.

Such principles are usually expressed more in word than in deed. I have sat through panels reviewing funding applications on global-health or medical research in Africa that did not include anyone from an affected country. Even when local experts are given appropriate responsibilities, anti-corruption and documentation requirements often go far beyond what is expected in funders' home countries. The obviously low expectations of African scientists and research institutions are as big a burden as the extra bureaucracy. The paradigm is beginning to shift, but the pace of this change needs to accelerate.

The four principles also apply to building institutional capacities. Most funders impose their own approaches to issues such as ethical review, financial management and data security. Finding out what systems are already in place and strengthening research institutions as needed would be more efficient. The Good Financial Grant Practice tool, launched last year by the African Academy of Sciences, is an exemplar of productive, respectful partnering.

Going forwards, the international community should focus on what African-led research offers: a distinct culture structured around knowledge gaps and desired impact, which Western science often struggles with. This mindset naturally leads to collaboration and high-quality science. My work on schistosomiasis, for example, requires me to work with local scientists, government officials, village health workers, teachers, mothers and other caregivers.

African-led research also builds sustainability. Local support must transcend any one project or funding scheme. My work in Zimbabwe has been supported by several funders. I am grateful to them all, but recognize that continuity and the ultimate impact require local commitment. Fortunately, it is self-reinforcing. Our collaborative approach has brought invitations to contribute to strategic initiatives such as the formulation of the African Union's Health Research and Innovation Strategy (HRISA 2018–2030), to be launched this month, and roadmaps for strengthening national health systems and continental vaccine policy.

Locally led partnerships are essential to producing relevant knowledge and sustainable change. The health of Africa, and the world, depends on making these happen.

News in brief

ATTACKS ON ACADEMIC FREEDOM WORLDWIDE RAISE CONCERN

Attacks on higher-education communities have become a troubling global phenomenon that shows no sign of abating, according to a report published on 19 November.

The annual analysis by international advocacy network Scholars at Risk, in New York City, tracks incidents that violate the academic freedom or human rights of scholars and students. This year, it documented 324 verified attacks in 56 countries from September 2018 to the end of August 2019, reported by volunteers (see 'Threatened freedom').

The incidents include violent or fatal attacks on scholars and students, wrongful prosecution or imprisonment, sacking or expulsion from institutions, travel restrictions, and systemic issues including university closures or military occupation of campuses.

This year's figure is up on the 294 incidents in 47 countries reported in 2018 – but the authors caution that more attacks are likely to go unreported.

Many of the most serious incidents occurred in countries struggling to develop effective academic communities. In five nations, reported attacks surged.

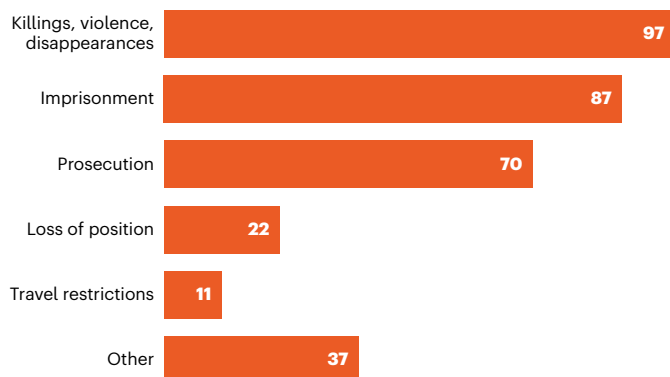
India has seen frequent violent confrontations between students and other groups on university campuses over issues including corruption and sexual harassment. Many clashes are violently suppressed by police.

And in the past year, Sudan has seen crackdowns by police and paramilitary organizations on political protests at higher-education institutions, amid political turmoil. This led to the temporary shutdown of many universities. The report also notes spikes in attacks on scholars and students in China, Brazil and Turkey.

The analysis comes as eight conservation researchers accused of spying were sentenced to prison in Iran last week. The researchers were arrested last year while studying the endangered Asiatic cheetah (*Acinonyx jubatus venaticus*). Authorities accused them of using camera traps to spy on sensitive infrastructure, but conservation and human-rights groups worldwide condemned the arrests and the subsequent trials in the country's revolutionary-court system. The researchers' prison sentences range from four to ten years (see page 566).

THREATENED FREEDOM

From 1 September 2018 to 31 August 2019, the Scholars at Risk network recorded 324 attacks that violated the academic freedom or human rights of scholars and students.



MODIFIED MOSQUITOES REDUCE CASES OF DENGUE

Disease-carrying mosquitoes are on the defensive. Cases of dengue fever, which is transmitted by the insects, plummeted in areas of Indonesia, Vietnam and Brazil in the months after researchers released *Aedes aegypti* mosquitoes that were modified to be resistant to dengue virus.

The findings, presented on 21 November at a meeting of the American Society of Tropical Medicine and Hygiene in National Harbor, Maryland, come from experimental releases of mosquitoes that carry *Wolbachia* bacteria (pictured), which block the replication of mosquito-borne pathogens such as the dengue and Zika viruses. The *Wolbachia* infection then spreads through local mosquito populations. The World Mosquito Program led the efforts, which it hopes will eventually be able to prevent mosquito-borne diseases.

Releases of *Wolbachia*-infected mosquitoes in 2016 near Yogyakarta City, Indonesia, led to a 76% reduction in cases of dengue fever over 2.5 years, compared with rates in areas where mosquitoes were not released. Two parts of Niterói, Brazil – home to around 500,000 people – experienced similar drops in dengue after releases in 2018. And declines occurred after a smaller 2018 release near Nha Trang, Vietnam.



Australia to resupply French Antarctic bases



L'Astrolabe, a naval transport ship, was supposed to be taking researchers and supplies to French bases in Antarctica (see above) this summer. Instead, the ship is currently docked in Hobart, Australia, after the navy discovered a critical defect in the ship's propeller.

In response to a call for assistance from Jérôme Chappellaz, director of the French Polar Institute, the Australian Antarctic Division (AAD) has made its icebreaker the RSV *Aurora Australis* available to carry French expeditioners and cargo – including food, equipment and 250,000 litres of fuel – to the French stations Dumont d'Urville and Concordia. The stations were in danger of being left without means of obtaining supplies this summer. The AAD will also take passengers to the French stations by air.

“Without the support of the AAD, the maintenance and supply of our research stations as well as the conduct of French scientific projects would have been at risk,” Chappellaz said in a media statement.

Kim Ellis, the AAD's director, said that Australia was pleased to assist the French programme, and that the required change in shipping schedule was unlikely to significantly affect Australia's research projects.

HOW STEM-CELL THERAPY BOOSTS HEART FUNCTION

Stem-cell therapies used to treat damaged heart muscle prompt an immune response that improves the heart's function, a team working on mice has revealed. The researchers report in *Nature* this week that they've uncovered the mechanism that explains the boost in heart function after stem-cell therapy, and have discovered how to mimic this repair with a chemical.

Stem-cell therapies for damaged hearts have shown some short-term improvements in animals, but the effect in humans has been limited. At first, scientists theorized that the benefits in mice came from stem cells differentiating into beating heart muscle cells, but further studies showed that this transformation did not happen.

Jeffery Molkentin, a cardiovascular-biology researcher at the Cincinnati Children's Hospital Medical Center in Ohio, and his team found that stem-cell therapies trigger immune cells called macrophages that help to repair connective tissue in the damaged area of the heart, which improves its function (R. J. Vagnozzi *et al.* *Nature* <https://doi.org/10.1038/s41586-019-1802-2>; 2019). The team showed that this repair mechanism can also be achieved using a chemical called zymosan, which is known to elicit an immune response.



TRUMP'S PICK TO LEAD OCEANS AGENCY WITHDRAWS

Barry Myers, US President Donald Trump's choice to lead the National Oceanic and Atmospheric Administration (NOAA), has withdrawn from consideration.

In a statement on his decision on 21 November, Myers (**pictured**) cited health concerns; he recently underwent surgery and started chemotherapy. His nomination has languished in the US Senate for more than two years without a vote.

An attorney by training, Myers is the former chief executive of the commercial forecasting firm AccuWeather in State College, Pennsylvania. His brother Joel Myers is now the company's chief executive.

Opponents of Myers' nomination have raised concerns about his lack of scientific credentials, as well as conflicts of interest surrounding any decisions involving NOAA's National Weather Service. Myers stepped down from his post at AccuWeather in January, more than a year after Trump nominated him to lead NOAA.

He has advocated giving the private sector a larger role in providing weather services. AccuWeather has lobbied for legislation that would prevent the National Weather Service from competing with private firms that provide basic weather forecasts and other data to the public.

News in focus



Archaeologists open a wooden coffin near Luxor, Egypt, in October.

RARE MUMMIFIED LIONS BOOST EXCITING TROVE OF EGYPTOLOGY FINDS

Egypt unveils lion mummies hot on the heels of a stunning human-coffin find – but the country is keeping research on these artefacts to itself for now.

By Antoaneta Roussi

The unveiling of five ancient feline mummies, including at least two lion cubs, and a host of other artefacts from the Bubasteion necropolis in Saqqara, south of Cairo, has left Egyptologists buzzing. It is only the second time that lion mummies have been found; the first was reported in 2004. The latest announcement, on 23 November, follows last month's revelation that 30 sealed coffins and their

mummified human contents had been discovered at the Assasif necropolis near Luxor, some 500 kilometres south of Saqqara.

Each lion mummy in the latest trove is around 1 metre long and dates to ancient Egypt's 26th Dynasty (664–525 BC), according to a statement from the Supreme Council of Antiquities, part of the Egyptian antiquities ministry. The artefacts were found with a larger collection of animal mummies, along with wood and bronze statues of cats and crocodiles.

"This is an extremely exciting and important

find, as it sheds new light on the relationship between wild and dangerous animals and the ancient Egyptians," says Salima Ikram, an archaeologist at the American University in Cairo who is working with the Egyptian government to identify the discoveries.

"It is possible that this hints at areas where lions were kept within Egypt, which we have seen [depicted] in images," she says, adding that mummified lions of this age are very rare.

More has also been revealed about the content and significance of the human coffins



L: KHALED DESOUKI/AFP/GETTY; R: HAMADA ELASAM/AP/SHUTTERSTOCK

One of the mummified felines found in Saqqara, Egypt (left) and a computed-tomography scan of one of the lion-cub mummies.

at Luxor. This find was first announced on 15 October, and is understood to be the largest coffin discovery since 153 were found at Bab el-Gasus ('the door of the priests'), not far from Luxor, in 1891.

Some of the coffins have been scanned using computed tomography, because the mummies cannot be unwrapped. The scans show the remains of a man aged 50, a woman of 35 and child 8–10 years old. All three are well preserved, and the child wore two gold bracelets, a ministry spokesperson told *Nature*.

The 30 Luxor coffins were discovered in two rows, stacked one on top of the other, in a pit 1 metre below ground. "I've never seen a parallel for this," says Kathlyn Cooney, chair of the department of Near Eastern languages and cultures at the University of California, Los Angeles.

Righting a wrong

The painted wooden sarcophagi are of a type known as stola coffins, after a set of red straps depicted on them. The garment would have been worn by people connected to the Amun priesthood, one of ancient Egypt's centres of power, dating back to the tenth century BC.

Stola coffins have intricate designs, which include complicated religious iconography, as well as details of personal information about the deceased. Cooney says it is crucial that the mummies have been found in their coffins, "potentially correcting, if not reversing, a century of colonial academic wrongs by Egyptologists who separated coffins from mummies and didn't or wouldn't study the human remains properly".

"Those interested in coffins will be able to look at the type of wood, varnish and paints;

those studying ancient pathologies will be able to examine the mummies' health," Cooney says. "For somebody like me, who studies the lives of past people, it helps bring those lives back to life," she adds.

But such anticipation is tempered by the knowledge that researchers outside Egypt will not yet be allowed to work on the finds, because the government is restricting research access to Egyptian institutions for now.

When *Nature* asked whether international researchers could contribute to the study of the discoveries – in line with the practice at many museums and heritage research institutions around the world – antiquities minister Khaled El-Enany replied: "We won't do a call [for proposals] on this study."

The team that found the coffins and the animal mummies was led by Mostafa Waziri, secretary-general of the Supreme Council of Antiquities. Waziri is not ruling out international involvement in research later on. But he confirms that any work will be led by Egypt's own researchers (see Editorial, page 565).

Willeke Wendrich, chair of African cultural archaeology at the University of California, Los Angeles, who became president of the International Association of Egyptologists this year, hopes that those in charge will eventually allow researchers from other countries to access the finds. Another Egyptologist, who asked to remain anonymous, urged the government not to delay, saying, "Science benefits from a multiplicity of talents."

SCIENCE FUNDERS GAMBLE ON GRANT LOTTERIES

A growing number of research agencies are assigning money randomly.

By David Adam

Albert Einstein famously insisted that God does not play dice. But the Health Research Council of New Zealand does. The agency is one of a growing number of funders that award grants

partly through random selection. Earlier this year, for example, David Ackerley, a biologist at Victoria University of Wellington, received NZ\$150,000 (US\$96,000) to develop new ways to eliminate cells – after his number came up in the council's annual lottery.

"We didn't think the traditional process was



L: KHALED DESOUKI/AFP/GETTY; R: HAMADA ELRASAM/AP/SHUTTERSTOCK

One of the mummified felines found in Saqqara, Egypt (left) and a computed-tomography scan of one of the lion-cub mummies.

at Luxor. This find was first announced on 15 October, and is understood to be the largest coffin discovery since 153 were found at Bab el-Gasus ('the door of the priests'), not far from Luxor, in 1891.

Some of the coffins have been scanned using computed tomography, because the mummies cannot be unwrapped. The scans show the remains of a man aged 50, a woman of 35 and child 8–10 years old. All three are well preserved, and the child wore two gold bracelets, a ministry spokesperson told *Nature*.

The 30 Luxor coffins were discovered in two rows, stacked one on top of the other, in a pit 1 metre below ground. "I've never seen a parallel for this," says Kathlyn Cooney, chair of the department of Near Eastern languages and cultures at the University of California, Los Angeles.

Righting a wrong

The painted wooden sarcophagi are of a type known as stola coffins, after a set of red straps depicted on them. The garment would have been worn by people connected to the Amun priesthood, one of ancient Egypt's centres of power, dating back to the tenth century BC.

Stola coffins have intricate designs, which include complicated religious iconography, as well as details of personal information about the deceased. Cooney says it is crucial that the mummies have been found in their coffins, "potentially correcting, if not reversing, a century of colonial academic wrongs by Egyptologists who separated coffins from mummies and didn't or wouldn't study the human remains properly".

"Those interested in coffins will be able to look at the type of wood, varnish and paints;

those studying ancient pathologies will be able to examine the mummies' health," Cooney says. "For somebody like me, who studies the lives of past people, it helps bring those lives back to life," she adds.

But such anticipation is tempered by the knowledge that researchers outside Egypt will not yet be allowed to work on the finds, because the government is restricting research access to Egyptian institutions for now.

When *Nature* asked whether international researchers could contribute to the study of the discoveries – in line with the practice at many museums and heritage research institutions around the world – antiquities minister Khaled El-Enany replied: "We won't do a call [for proposals] on this study."

The team that found the coffins and the animal mummies was led by Mostafa Waziri, secretary-general of the Supreme Council of Antiquities. Waziri is not ruling out international involvement in research later on. But he confirms that any work will be led by Egypt's own researchers (see Editorial, page 565).

Willeke Wendrich, chair of African cultural archaeology at the University of California, Los Angeles, who became president of the International Association of Egyptologists this year, hopes that those in charge will eventually allow researchers from other countries to access the finds. Another Egyptologist, who asked to remain anonymous, urged the government not to delay, saying, "Science benefits from a multiplicity of talents."

SCIENCE FUNDERS GAMBLE ON GRANT LOTTERIES

A growing number of research agencies are assigning money randomly.

By David Adam

Albert Einstein famously insisted that God does not play dice. But the Health Research Council of New Zealand does. The agency is one of a growing number of funders that award grants

partly through random selection. Earlier this year, for example, David Ackerley, a biologist at Victoria University of Wellington, received NZ\$150,000 (US\$96,000) to develop new ways to eliminate cells – after his number came up in the council's annual lottery.

"We didn't think the traditional process was

appropriate,” says Lucy Pomeroy, the senior research-investment manager for the fund, which began its lottery in 2015. The council was launching a new type of grant, she says, aiming to fund transformative research, so wanted to try something new to encourage fresh ideas.

Traditionalists beware: the forces of randomness in research are, if not quite on the march, then certainly plotting their next move. At a meeting at the University of Zurich in Switzerland on 19 November, supporters of the approach argued that blind chance should have a greater role in the scientific system.

And they have more than just grant applications in their sights. They say lotteries could be used to help select which papers to publish – and even who to appoint to academic jobs.

Luck of the draw

“Random chance will create more openness to ideas that are not in the mainstream,” says Margit Osterloh, an economist at the University of Zurich who studies research governance and organized the meeting, which was intended to promote the idea among academics. She says that existing selection processes are inefficient. Scientists have to prepare lengthy applications, many of which are never funded, and assessment panels spend most of their time sorting out the specific order in which to place mid-ranking ideas.

Low- and high-quality applications are easy to rank, she says. “But most applications are in the midfield, which is very big.” Most importantly, she argues, standard assessments don’t perform as well as policymakers, publishers and university officials assume. “Referees and all kinds of evaluation bodies do not have really good working criteria.”

The Swiss National Science Foundation (SNSF) is the latest funder to experiment with random selection. Earlier this year, it asked assessment panels to draw lots to help decide which early-career scientists should receive postdoctoral fellowships. It is now evaluating the scheme, and SNSF president Matthias Egger spoke about it at the Zurich meeting. Other programmes that rely on lottery systems to award some grant types include another New Zealand government fund called the Science for Technological Innovation National Science Challenge (SfTI), which introduced random selection in 2015. Germany’s largest private funding agency, the Volkswagen Foundation in Hanover, has also used lotteries to allocate some of its Experiment! grants since 2017.

‘We actually do have a hat’

The process is not entirely random. Typically, funders screen applications to ensure that they meet a minimum standard, then projects are given numbers and selected at random by a computer until all the cash has been allocated.



Lotteries are increasingly being used to choose which grant applications should get money.

“It just takes a lot of angst out of it,” says Don Cleland, a process engineer at Massey University in Palmerston North, New Zealand, and a member of the team that oversees the SfTI fund.

Given the money to fund 20 projects, an assessment panel doesn’t need to agonize over which application ranks 20th and which comes 21st, he says. The panel members can just agree

“Random chance will create more openness to ideas that are not in the mainstream.”

that both are good enough to be funded, and then put them into the hat. “We actually do have a hat,” Cleland says.

The fund tells applicants how far they got in the process, and feedback has been positive, he says. “Those that got into the ballot and miss out don’t feel as disappointed. They know they were good enough to get funded and take it as the luck of the draw.”

The idea has some theoretical backing. A number of researchers have analysed various selection methods and suggested that incorporating randomness has advantages over the current system, such as reducing the bias that research routinely shows plagues grant-giving, and improving diversity among grantees (F. C. Fang & A. Casadevall *mBio* 7, e00422-16; 2016).

The acceptance criteria for entering the lottery can be tweaked, for example, to give more weighting to scientists from minority ethnic backgrounds or to those who aren’t backed

by wealthy institutions. People from wealthy institutions or privileged backgrounds often have access to resources that help them to achieve success by standard metrics. And the conventional system tends to benefit them, says Cleland, because it focuses on candidates’ track records rather than the strength of their ideas. “We want those with the best ideas to rise to the top.”

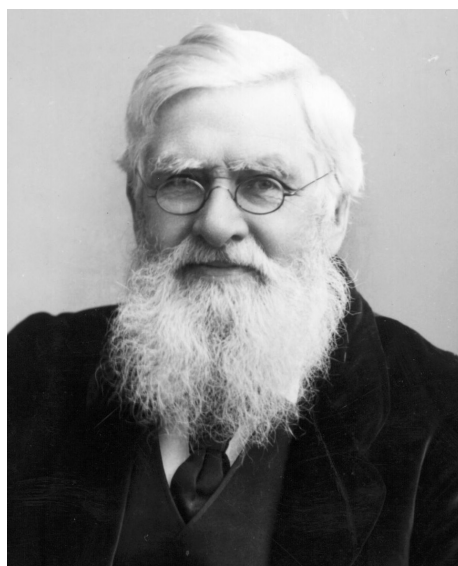
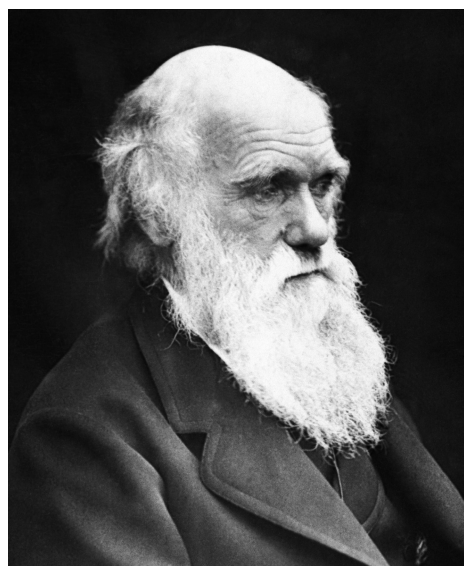
Competitive arguments

Cleland argues that other funders should try it. But not everyone agrees. Despite benefiting from a grant lottery, Ackerley says he doesn’t approve of them. “I spend a lot of time on grant-review panels and I like to think they do a reasonable job,” he says. “I’ve done reasonably well out of competitive grants and I suppose the selfish reason is that I might not do so well out of a lottery system.”

Because applications to funds that use lottery systems only need to satisfy basic criteria, they tend to be shorter. “I think there’s a lot of value to writing a high-quality proposal,” Ackerley says.

Osterloh, who triggered lively debate of her arguments in the pages of *Research Policy* after publishing them in the journal earlier this year (M. Osterloh & B. S. Frey *Res. Policy* 49, 103831; 2020), says selection by random chance could have a wider advantage because those who benefit from lotteries do not feel so entitled.

“If you know you have got a grant or a publication which is selected partly randomly, then you will know very well you are not the king of the Universe, which makes you more humble,” she says. “This is exactly what we need in science.”



Charles Darwin (left) rushed to publish *On the Origin of Species* after receiving a manuscript detailing similar ideas from Alfred Russel Wallace (right).

SCOOPED IN SCIENCE? RELAX, CREDIT WILL COME YOUR WAY

A study of races to solve protein structures shows that teams that come second still get recognition.

By Ewen Callaway

Being scooped to a discovery is a scientist's worst nightmare. But the penalties for coming second aren't as harsh as some might think.

Scooped papers receive only about one-quarter fewer citations than do papers that were the first to report the same discovery, according to an analysis of more than 1,600 'races' to determine the detailed 3D shape, or structure, of proteins and other biomolecules.

"You get a meaningful advantage for being first, but being scooped may not be as devastating as people seem to fear," says Carolyn Stein, an economist at the Massachusetts Institute of Technology (MIT) in Cambridge who conducted the study with her MIT colleague Ryan Hill, also an economist. Their results are described in a working paper posted to the MIT website (see go.nature.com/2xmcuyv).

Social scientists say that the research breaks new ground because it is able to identify and track scooped studies, including some that were never published – although they caution that the findings might not apply to other fields. "This is the first study I'm aware of that has been able to observe unpublished papers," says Michaël Bikard, an innovations researcher at

the French campus of business school INSEAD in Fontainebleau. "This is important stuff."

The history of science is rife with competition. Charles Darwin rushed out his *On the Origin of Species* after receiving a manuscript detailing similar ideas from Alfred Russel Wallace – and Isaac Newton, Gottfried Wilhelm Leibniz and their supporters feuded over who invented calculus.

Despite the prominence of such rivalries, scholars of science know little about how credit is apportioned for competing discoveries. Theoretical models analysing patent races, for instance, often assume that to the victor go all the spoils. In the real world, however, credit for scientific discoveries is unlikely to be winner-takes-all, say researchers.

Protein probe

One problem with studying scooped projects is that some scientists abandon a research effort after someone else has beaten them to it, says Hill, a PhD student who was inspired to do the study partly as a result of being scooped during his graduate work. Alternatively, researchers modify the project in such a way that it is impossible to compare its results with those of the paper that scooped it.

In search of an 'apples-with-apples' comparison of competing projects, Hill and Stein

used the Protein Data Bank (PDB), a repository of more than 150,000 structures of proteins and other biomolecules. These structures are key to understanding how proteins work, as well as how their function might be altered by drugs. Crucially for the study, scientists tend to submit structures to the PDB – under embargo – months before a paper describing the work is published in a journal (and the embargo on the PDB structure is lifted). This approach allowed the researchers to follow 1,630 'races' in which competing teams submitted to the PDB structures of the same, or closely similar, molecules between 1999 and 2017.

The cost of being scooped was moderate. Structures released second were only 2.5% less likely ever to be published, although they tended to appear in less prestigious journals (as measured by impact factor), than were structures published first. Hill and Stein estimate that, as a share of total citations out of a 100, the first paper would receive 58 and the second paper 42.

But when questioned about the effects of being scooped, scientists were much more pessimistic than those data show, according to Hill and Stein's survey of 915 structural biologists. The scientists overestimated the odds of being beaten to a discovery, and predicted that, out of 100 citations, a scooped paper would receive just 29.

But not all scientists were penalized equally for coming second, the study found. When research teams at leading universities and departments – as measured by a universities ranking table – were beaten by a team at a lower-profile institution, the second-placed team got slightly more citations. And the teams at top institutions accrued an even larger share of citations when they did the scooping.

"I was blown away by this result," says Bikard.

Race for recognition

Paula Stephan, an economist at Georgia State University in Atlanta, says the study is the first she knows of that actually measured the penalty for being scooped. "We have known for many years that science is not a winner-takes-all 'game'. This piece of research confirms this." But she cautions against generalizing the study to other fields.

And the study doesn't capture the psychological effects of being scooped, says Venki Ramakrishnan, a structural biologist at the Laboratory of Molecular Biology in Cambridge, UK. In the late 1990s and early 2000s, his group raced several teams to determine the structure of the ribosome, a cellular machine that makes proteins. In September 2000, a team led by Ada Yonath at the Weizmann Institute in Rehovot, Israel, published the structure of a ribosome subunit in *Cell*¹ that Ramakrishnan's team had also characterized. Ramakrishnan's study came out weeks later in *Nature*².

"For that month, I and my lab were pretty

L: BETTMANN/GETTY; R: LONDON STEREO SCOPIC COMPANY/GETTY

miserable,” he says. The researchers worried that they wouldn’t receive proper recognition for their work. That didn’t turn out to be the case. Ramakrishnan’s and Yonath’s teams are both credited with working out the ribosome-subunit structure – and the scientists each

shared one-third of the 2009 Nobel chemistry prize. Ramakrishnan’s paper has roughly twice as many citations as the one that scooped it. “In the long run, it didn’t matter,” he says.

1. Schluzen, F. et al. *Cell* **102**, 615–623 (2000).
2. Wimberly, B. T. et al. *Nature* **407**, 327–339 (2000).

GLOBAL 5G WIRELESS DEAL THREATENS WEATHER FORECASTS

Meteorologists say lax international standards could degrade crucial satellite measurements.

By Alexandra Witze

The international agency that regulates global telecommunications agreed to new radio-frequency standards on 21 November. Meteorologists say the long-awaited decision threatens the future of weather forecasting worldwide by allowing transmissions from mobile-phone networks to degrade the quality of Earth observations from space.

Wireless companies are beginning to roll out their next-generation networks, known as 5G, around the world. The new agreement is meant to designate the radio frequencies over which 5G equipment can transmit. But some of the frequencies come perilously close to those used by satellites to gather crucial weather and climate data. To keep the signals from interfering with one another, researchers

have proposed turning down the amount of noise allowed to leak from 5G transmissions.

Negotiators at a meeting of the International Telecommunication Union in Sharm El-Sheikh, Egypt, agreed to introduce two stages of protection for frequencies near 24 gigahertz – a range close to those that weather satellites use to detect the amount of water in the atmosphere. Companies that operate 5G networks will have a relatively loose standard from now until 2027. After that, the regulation will get stricter. The idea is to let 5G companies start building networks now, and then to add more protection for Earth observations as 5G transmissions become denser.

But having eight years with relatively lax regulation is “of grave concern” to weather forecasters, says Eric Allaix, a meteorologist at Météo-France in Toulouse who heads a World Meteorological Organization (WMO) group on

radio-frequency coordination. The WMO is so upset that it included a statement of concern in the meeting minutes, he says.

“The race for 5G is going to go fast,” says Renée Leduc, a consultant with Narayan Strategy in Washington DC who works on spectrum-sharing issues. “In the early-to-mid-2020s we’re going to see a very quick uptick.” Although more protections for Earth observations will take effect in 2027, “I’m still really concerned about the time period between now and then,” she says.

The 5G transmissions will involve many frequencies, but the key one under discussion is 23.8 gigahertz. Water vapour in the atmosphere naturally produces a weak signal at this frequency, which satellites use to measure humidity. Those data feed into weather forecasts. But if a 5G station is transmitting a signal near the 23.8-gigahertz frequency, a weather satellite might pick it up and mistakenly interpret it as water vapour.

Meteorologists say that the problem is manageable, but only if there is enough of a noise buffer between 5G transmissions and the water-vapour signal. The buffer, measured in decibel watts, is akin to a gauge of how much you might turn down the volume of your stereo so as not to bother your neighbours.

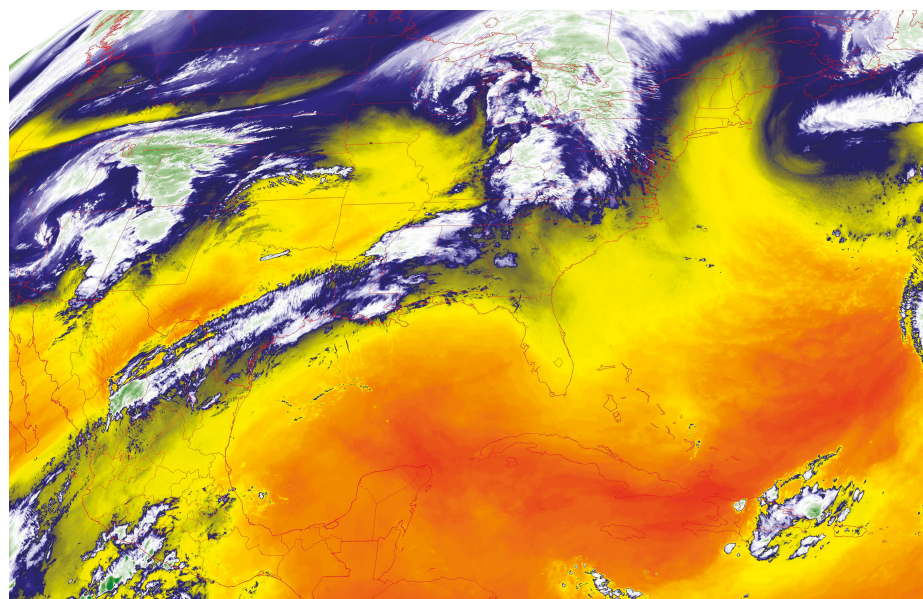
In the run-up to the Egypt conference, the WMO had been pushing for a buffer of –55 decibel watts. European regulators had settled on a less-stringent recommendation of –42 decibel watts for 5G base stations. The US Federal Communications Commission had advocated just –20 decibel watts.

The new standard hews closest to the European proposal: it is –33 decibel watts until September 2027, and –39 decibel watts after that.

“These two values were set by long negotiations between the member states,” said David Botha, a counsellor with the International Telecommunication Union, at a 22 November media briefing. “These values were considered to be adequate, in the sense that they would provide protection to the weather satellite systems, to Earth-exploration satellite systems. We have nevertheless noted that there were concerns that were issued.”

Even the stricter level is not enough to avoid interfering with water-vapour measurements, says Leduc. A US government study found that 5G base stations needed to transmit with a noise buffer of –52.4 decibel watts to protect the water-vapour observations.

Weather forecasters will have to gauge how to mitigate the impacts on satellite observations – perhaps by working with the wireless industry to research ways to shut down or redirect 5G transmissions when a satellite is making its measurements. Botha said that the agreement requires a “continued monitoring” of how 5G networks affect weather observations, but he provided no details on what that would involve.



Water vapour over the Americas is shown in this US government satellite image.

miserable,” he says. The researchers worried that they wouldn’t receive proper recognition for their work. That didn’t turn out to be the case. Ramakrishnan’s and Yonath’s teams are both credited with working out the ribosome-subunit structure – and the scientists each

shared one-third of the 2009 Nobel chemistry prize. Ramakrishnan’s paper has roughly twice as many citations as the one that scooped it. “In the long run, it didn’t matter,” he says.

1. Schlutzen, F. et al. *Cell* **102**, 615–623 (2000).
2. Wimberly, B. T. et al. *Nature* **407**, 327–339 (2000).

GLOBAL 5G WIRELESS DEAL THREATENS WEATHER FORECASTS

Meteorologists say lax international standards could degrade crucial satellite measurements.

By Alexandra Witze

The international agency that regulates global telecommunications agreed to new radio-frequency standards on 21 November. Meteorologists say the long-awaited decision threatens the future of weather forecasting worldwide by allowing transmissions from mobile-phone networks to degrade the quality of Earth observations from space.

Wireless companies are beginning to roll out their next-generation networks, known as 5G, around the world. The new agreement is meant to designate the radio frequencies over which 5G equipment can transmit. But some of the frequencies come perilously close to those used by satellites to gather crucial weather and climate data. To keep the signals from interfering with one another, researchers

have proposed turning down the amount of noise allowed to leak from 5G transmissions.

Negotiators at a meeting of the International Telecommunication Union in Sharm El-Sheikh, Egypt, agreed to introduce two stages of protection for frequencies near 24 gigahertz – a range close to those that weather satellites use to detect the amount of water in the atmosphere. Companies that operate 5G networks will have a relatively loose standard from now until 2027. After that, the regulation will get stricter. The idea is to let 5G companies start building networks now, and then to add more protection for Earth observations as 5G transmissions become denser.

But having eight years with relatively lax regulation is “of grave concern” to weather forecasters, says Eric Allaix, a meteorologist at Météo-France in Toulouse who heads a World Meteorological Organization (WMO) group on

radio-frequency coordination. The WMO is so upset that it included a statement of concern in the meeting minutes, he says.

“The race for 5G is going to go fast,” says Renée Leduc, a consultant with Narayan Strategy in Washington DC who works on spectrum-sharing issues. “In the early-to-mid-2020s we’re going to see a very quick uptick.” Although more protections for Earth observations will take effect in 2027, “I’m still really concerned about the time period between now and then,” she says.

The 5G transmissions will involve many frequencies, but the key one under discussion is 23.8 gigahertz. Water vapour in the atmosphere naturally produces a weak signal at this frequency, which satellites use to measure humidity. Those data feed into weather forecasts. But if a 5G station is transmitting a signal near the 23.8-gigahertz frequency, a weather satellite might pick it up and mistakenly interpret it as water vapour.

Meteorologists say that the problem is manageable, but only if there is enough of a noise buffer between 5G transmissions and the water-vapour signal. The buffer, measured in decibel watts, is akin to a gauge of how much you might turn down the volume of your stereo so as not to bother your neighbours.

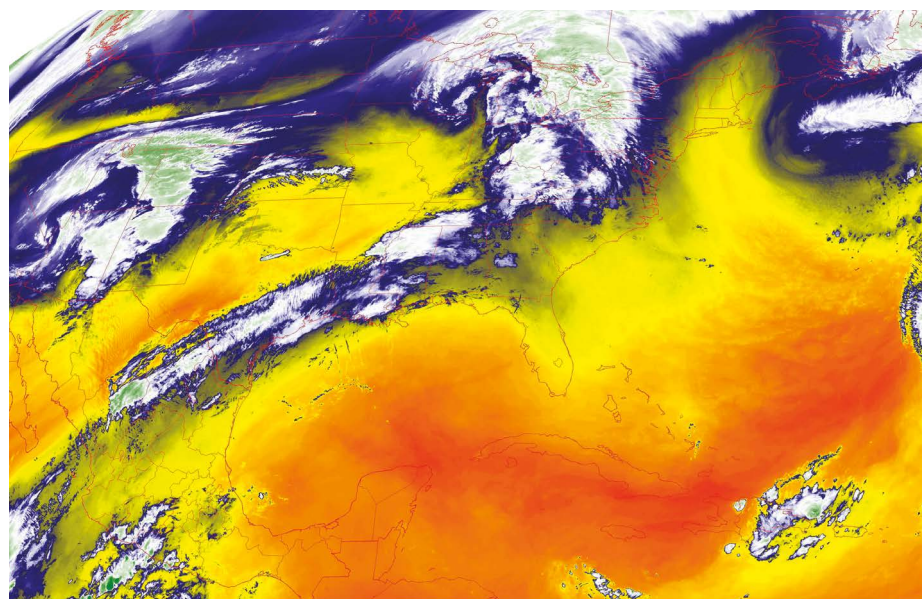
In the run-up to the Egypt conference, the WMO had been pushing for a buffer of –55 decibel watts. European regulators had settled on a less-stringent recommendation of –42 decibel watts for 5G base stations. The US Federal Communications Commission had advocated just –20 decibel watts.

The new standard hews closest to the European proposal: it is –33 decibel watts until September 2027, and –39 decibel watts after that.

“These two values were set by long negotiations between the member states,” said David Botha, a counsellor with the International Telecommunication Union, at a 22 November media briefing. “These values were considered to be adequate, in the sense that they would provide protection to the weather satellite systems, to Earth-exploration satellite systems. We have nevertheless noted that there were concerns that were issued.”

Even the stricter level is not enough to avoid interfering with water-vapour measurements, says Leduc. A US government study found that 5G base stations needed to transmit with a noise buffer of –52.4 decibel watts to protect the water-vapour observations.

Weather forecasters will have to gauge how to mitigate the impacts on satellite observations – perhaps by working with the wireless industry to research ways to shut down or redirect 5G transmissions when a satellite is making its measurements. Botha said that the agreement requires a “continued monitoring” of how 5G networks affect weather observations, but he provided no details on what that would involve.



Water vapour over the Americas is shown in this US government satellite image.

CHINESE INFILTRATION OF US LABS CAUGHT AGENCIES OFF GUARD, REPORT SAYS

China has diverted US government funds to bolster its military and economic aims, a US Senate panel finds.

By Jeff Tollefson

US science agencies' slow response to the threat posed by China's talent-recruitment programmes has allowed China to divert US government funds and private-sector technology to further its military and economic goals, a US Senate panel has found.

Its report, which lawmakers discussed at a hearing on 19 November, describes new details of what it says are China's efforts to infiltrate US research institutions – including contract provisions requiring participating scientists to work on behalf of China. The analysis focused on China's Thousand Talents Plan, the most prestigious of more than 200 programmes that are designed to recruit leading academics and promote domestic research.

Federal science agencies have been caught off guard by these programmes, lawmakers said, and they must now coordinate efforts to protect the US research enterprise. "We have to be nimble," said Senator Rob Portman (Republican, Ohio).

He pointed to provisions in sample Thousand Talents contracts that require participating scientists to keep the contract secret, recruit postdocs and sign over any intellectual-property rights to the sponsoring Chinese institution. The contracts provide incentives for scientists to set up 'shadow labs' in China that mirror US taxpayer-funded research, the Senate report says.

Michael Lauer, a deputy director at the US National Institutes of Health (NIH), told lawmakers that those laboratories allow China to see what is happening in the United States

before the rest of the world. When the NIH informed US research institutions about these shadow labs as part of a broader investigation into foreign influence, confidentiality rules and conflicts of interest, the news often came as a surprise, Lauer said.

The report also includes examples of potential undue foreign influence at the US National Science Foundation and the departments of commerce, energy and state. One post-doctoral researcher at an energy-department lab who was also part of China's Thousand Talents Plan removed 30,000 electronic files from the US lab before returning to China, the report says.

Probes into the issue by the NIH and other government agencies have fuelled fears that researchers of Chinese descent are being targeted unfairly. They have also left some institutes struggling to balance security concerns with academic openness.

By delving into requirements in talent-programme contracts, the Senate report provides a vivid depiction of these issues, says Tobin Smith, vice-president for policy at the Association of American Universities in Washington DC.

"This will help us as we try to make faculty aware of why they ought to be careful in entering into any of these talent programmes," Smith says.



CellSorter

COMPACT DEVICE FOR
**ROBOTIC SINGLE
CELL ISOLATION**

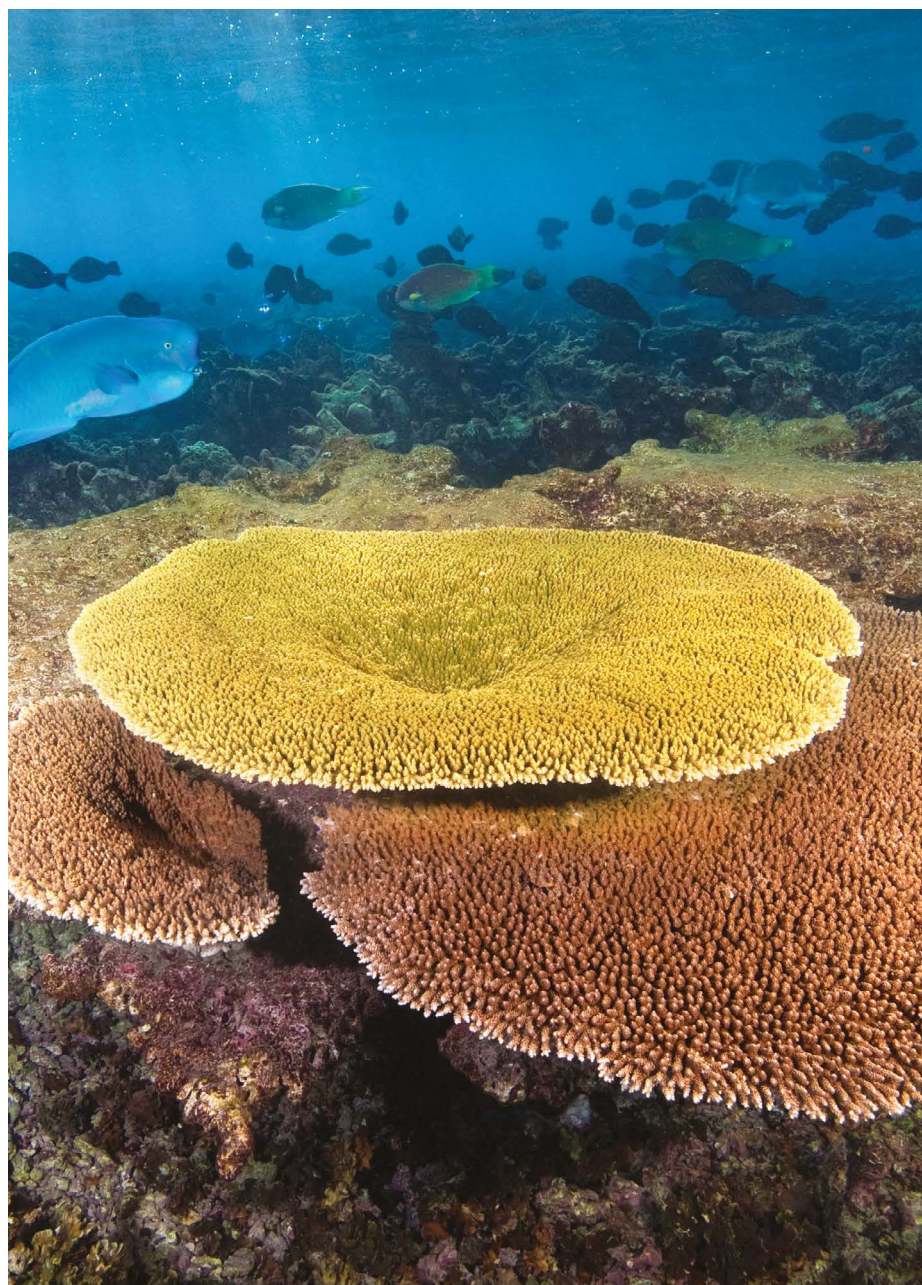
High-throughput screening and computer
vision for rare and sensitive cells...

WWW.SINGLECELLPICKER.COM




NATIONAL RESEARCH, DEVELOPMENT
AND INNOVATION OFFICE
HUNGARY

PROJECT
FINANCED FROM
THE NRDI FUND
MOMENTUM OF INNOVATION



Some coral reefs off the Phoenix Islands in Kiribati seem to be resilient to warming seas.

HOPE FOR CORAL REEFS

The ocean is warming and reefs are fading. But optimistic marine scientists are working to keep some corals alive until the climate stabilizes. **By Amber Dance**

Anne Cohen dropped into the ocean off the coast of the Phoenix Islands expecting to find desolation. It was 2018, and a powerful El Niño weather system two years earlier had warmed the waters around this mid-Pacific atoll by nearly 3 °C. Coral reefs simmered in the heat.

Such feverish temperatures cause the tiny animals that make up a reef to expel the colourful, symbiotic algae that nourish them. They bleach, starve and die. On her expedition to the islands, part of the nation of Kiribati, Cohen found greyish reefs in which almost 70% of corals had expired.

But she also found reason for hope.

“We’d come across these areas, I’m talking about several square kilometres, with super-high coral cover and super-high coral diversity,” recalls Cohen, a marine scientist at the Woods Hole Oceanographic Institution in Massachusetts. Healthy taupe coral branches sprouted from a field of blonde and rose plates, while schools of gold-and-magenta anthias fish flitted to and fro.

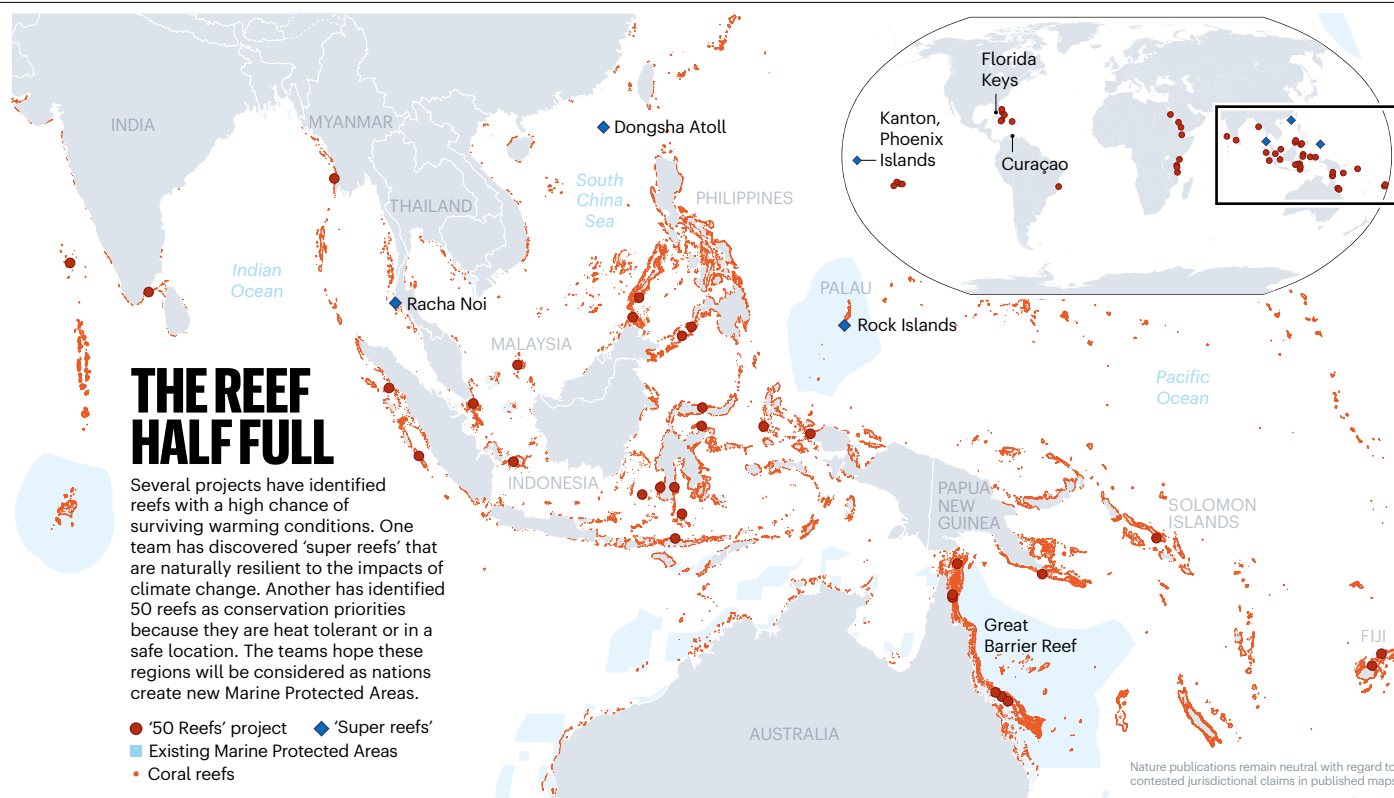
Such places give ocean ecologists hope that even as the climate warms, corals still have a fighting chance. When these scientists hear that 70–90% of reefs could be gone by mid-century, they focus on the 30% that might live. And they’re taking action to save those reefs for the future. Around the world, hundreds of millions of people depend on reefs for food, tourism income and protection from ocean storms.

Researchers and conservationists are testing many strategies to help corals. Some approaches – such as bredisecding or genetically manipulating corals to tolerate climate change, or sprinkling reefs with beneficial microbes¹ – have yet to prove themselves outside of the laboratory.

But scientists are already trying out other ideas in the wild: growing and replanting corals in damaged reefs, for instance, and helping them to breed. And there’s another tactic getting attention – finding the reefs that have the best natural chances of survival and helping them to stay alive. There are downsides to both of these approaches, however: breeding is impractical at large scales, and some researchers worry that focusing only on naturally resilient reefs will constrain conservation to a few niche locations.

Conservationists say that it will take a variety of strategies to save coral reefs, but the time is now. “The next decade is really our window,” says Lizzie McLeod, who works in reef management at the Nature Conservancy in Alexandria, Virginia.

Half the globe’s reefs are already gone. Carbon pollution heats up the ocean and turns it more acidic, making it difficult for the delicate creatures to build their calcium carbonate skeletons. Human activities such as fishing, dredging, pollution and development have



also caused serious harm.

The efforts to save particular reefs are offering some hope, researchers say, but these solutions are only a stop-gap. "Unless we curb carbon emissions, none of this is going to make any difference whatsoever," says Iliana Baums, a molecular ecologist at Pennsylvania State University (Penn State) in University Park.

Helping hands

In the Caribbean Sea, a one-two punch of climate change and disease has hit reefs hard. "The Florida reef tract has been going downhill for the last 40 years or so," says Erinn Muller, science director of the Elizabeth Moore International Center for Coral Reef Research & Restoration at the Mote Marine Laboratory & Aquarium in Sarasota, Florida. "We really haven't seen any natural recovery."

So conservationists from Mote and other organizations are taking matters into their own hands. Over the past several years, the Mote laboratory's reef-restoration teams have planted almost 70,000 pieces of coral from five main species off the Florida Keys. The goal of this project, and others like it around the world, is not to replant reefs entirely, but to provide enough new corals to allow them to reproduce themselves.

Mote's coral nannies start with artificial, ocean-based nurseries. They create tree-like structures by hanging fishing line from buoys and attaching plastic branches. Then they dangle individual corals from those branches.

Once the dangles reach about the size of a basketball, divers carve out softball-sized chunks. Then Muller and her colleagues

test the corals, ensuring that some in each batch have the genes to resist dangers such as diseases, heat or acidification, before the restoration crew plants them on the sea floor.

The reef begins to recover within a year, says Muller, and more fish and invertebrates move in. Eventually, the planted chunks will fuse into larger corals. From their experiences with past plantings, the researchers know that the corals will be big enough to spawn within a couple of years.

This works for fast-growing, branching corals. But some mound and boulder corals expand slowly — just a couple of millimetres per year. Mote researchers speed that up by cutting the corals into tiny pieces and affixing them to small, round tiles with stems, like a ceramic flower². In the lab's on-shore tanks, the finger-nail-sized fragments grow up to 50 times faster than the full-sized corals, reaching 3 centimetres across in a few months.

To replant them, the restoration team finds a large coral head from which the animals have disappeared but left their limestone skeletons. Divers drill about 20 holes in that dead coral, and poke in the stems of the coral-tile flowers. Eventually, the pieces grow together, 're-skinning' the dead coral and restoring it to spawning in two to three years.

"It's a way to jump-start the reef," says Emily Hall, a chemical ecologist at Mote. The lab places thousands of these coral-flowers every year, says Muller, and typically achieves more than 80% survival.

The microfragmentation approach is promising, says Dirk Petersen, executive director of the coral conservation non-profit organization

SCORE International in Bremen, Germany. But he has doubts about re-skinning. The old coral head might have new residents that could harm or compete with the coral-flowers, he says, and organisms that killed the original inhabitants might still be present.

Romancing the reef

Mote, SCORE and others are also helping corals to reproduce. By assisting breeding, reef managers can make not just more corals, but more diverse populations. Some larvae will be heat-tolerant, some resistant to acidified waters. Nature will select those that match a given reef's conditions.

It was in pursuit of this diversity that Baums went to the Caribbean island of Curaçao in August. She and her colleagues from SCORE teamed up with local conservationists to perform *in vitro* fertilization for elkhorn coral (*Acropora palmata*).

The affair began a couple of hours after sunset, under a full moon. Corals spewed their gametes into the water. Sperm and eggs were so thick, they formed a slick on the surface. Baums and the other divers secured nets over the colonies to collect the gametes.

Because gametes are viable only for a few hours, the team had to race to pair up compatible eggs and sperm. "We have used a bench on the beach as an improvised laboratory," says Petersen. The low-tech lab can boost fertilization rates to 90% or higher.

Then, SCORE placed the embryos in the nursery, a floating pool in the sea that is protected from predators. A few days later, larvae settled on 3D-printed, starfish-shaped

structures. When the corals reach about fingernail size, divers will wedge these substrates into the crevices of needy reefs. Petersen says this approach has also been tested in reefs off Mexico, Florida, the Bahamas, the Dominican Republic, Bonaire, Australia, Palau and Guam.

Although these restoration measures are beneficial in deeply damaged reefs, they are difficult and costly to deploy at scale. Consider Australia's Great Barrier Reef, which is close to the size of Italy. Heatwaves in 2016 and 2017 bleached half the corals. Although some restoration is happening, surviving corals in larger areas have to help themselves, says Terry Hughes, director of the Australian Research Council Centre of Excellence for Coral Reef Studies at James Cook University in Townsville. "Billions of tougher corals survived, and they're breeding."

Come hell or hot water

Fortunately, the Great Barrier Reef's neighbourhood is better off than the Caribbean's. "The Indo-Pacific is still in great shape," says Emily Darling, a conservation scientist at the Wildlife Conservation Society in Toronto, Canada.

In August, a large team led by Darling published an analysis of 2,584 Indo-Pacific reefs³. The good news is that 86% of those reefs were dominated by large species such as branching, plating and boulder corals, the types that create fish habitats and shield shorelines from storms.

The bad news is that these species are also the most sensitive to heat.

Darling and her colleagues identified a group of almost 450 reefs that had been affected very little by recent warming events and had retained more than 10% coral cover – the minimum at which the reef can build more skeleton than it's likely to lose (but still well below the area's coral cover in decades past). "Right now, I would really target those to save," says Darling.

Cohen, too, is on a quest to find resilient reefs and make sure they are protected in her Super Reefs project (see 'The reef half full').

Her team has found three ways in which reefs beat the heat. Some corals live in naturally warm environments and are genetically adapted to deal with scorching temperatures. For example, reefs off Palau's Rock Islands withstood major heatwaves in 1998 and 2010. In the nearby barrier reef, where the water is typically cooler, coverage dropped to 5–6%. Other reefs are simply lucky in location. Some benefit from cold currents that protect them from hot spells⁴. Others are served by currents that provide a constant plankton buffet. Even if they bleach, these corals remain well-fed and survive.

"So far," says Cohen, "most of the super reefs we've found have been by accident." To find more, she's using hydrodynamic modelling of currents to identify likely spots, and developing underwater robots that can identify

areas with living corals and hover over them to gather data. "I'm convinced there are super reefs in most places," she says.

The next step will be to protect them. For example, Cohen is working with the Nature Conservancy and the government of the Marshall Islands to identify resilient reefs in the nearby waters, and to ensure that the nation includes some of these corals in any new marine protected areas.

Another researcher on the hunt for resilient reefs is Ove Hoegh-Guldberg, a marine scientist at the University of Queensland in Brisbane, Australia. He borrowed an algorithm from the world of finance to find them.

Investors want a diverse portfolio that balances risk and reward. Hoegh-Guldberg

"I'm convinced there are super reefs in most places."

joined forces with other coral specialists and economists to apply the same logic to the world's reefs. For example, being in the predicted path of cyclones might put a reef at high risk, but having good tolerance for climate change, or being well-connected to other reefs that could need re-seeding, would give it a high benefit profile.

Considering 30 such factors, the group identified 50 ocean regions scattered throughout the tropics, each about 500 square kilometres, as conservation priorities⁵. If the rest of the world's reefs are wiped out, these reefs might be able to kick-start repopulation, assuming the climate stabilizes, Hoegh-Guldberg says.

Some find the idea of singling out survivors troubling. "I hope it doesn't come to the point that we have to pick and choose one reef over another," says Hall. "I think all reefs are important."

Mónica Medina, a coral-reef biologist at Penn State, called the idea of relying on naturally robust corals "dangerous" at an October meeting of science writers held at her university. Corals that are resilient in one habitat might not thrive elsewhere, she pointed out, and the idea can lull people into thinking that reefs don't need other assistance.

Both Hoegh-Guldberg and Cohen stress that they don't intend for only these special reefs to receive conservation attention. In his '50 Reefs' project, Hoegh-Guldberg is simply hedging his bets by identifying some promising targets. "You have a good chance of winning, no matter what."

Social science

In one of his 50 regions, off Madagascar, the wildlife charity WWF is helping protect a beautiful reef. Broad table corals rise in layers, and sea turtles prowl the clear waters. Although there are some signs of bleaching from past stress,

the reef has mostly recovered, says Gabby Ahmadi, a marine conservation scientist with the WWF in Washington DC. But there aren't quite as many fish as there ought to be.

The way to solve that fish deficit is as much about local people as it is about marine ecology, she says. Conservation groups have helped locals to set up boutique tourism. The Nosy Hara marine national park pays community rangers to keep an eye on the corals. Fishers across Madagascar teach each other protective practices.

Conservationists are applying similar strategies in several places, customized to the people and the environment. It's societal action, not biology, that will ultimately save reefs, Hughes says. "They'll die again if we don't first fix water pollution, overfishing and ever-rising emissions."

Nonetheless, conservationists cling to hope for the future – albeit tempered with realism. "I think we can resolve this crisis," says Petersen. "I'm also sure that reefs will look different than they do now." Coral ecologists might have to sacrifice species diversity to rescue the whole ecosystem. Petersen compares future reefs to modern, managed forests, which host a narrower mix of trees than untouched forests, but still provide habitats for wildlife and help to cleanse water and air.

Another reason for hope is that reefs have made it through other challenges. Reef-building corals are part of a lineage that is more than 400 million years old. They've endured global water temperatures that have swung between 10 and 32 °C, and carbon dioxide levels up to quadruple those of today. But they've never before had to endure such rapid warming.

The current pace of climate change could stretch corals beyond their adaptive limits. Even if nations succeeded in limiting global warming to 1.5 °C above 1990 levels, the target of the 2018 Paris agreement, 70–90% of reefs would be lost, according to the Intergovernmental Panel on Climate Change. And that target is looking more and more unrealistic. If the climate warms by 2 °C, the panel projects⁶ losses of greater than 99%.

Still, researchers are confident that with human assistance, and a global drop in emissions, reefs can rise again. "Corals 'know' how to recover, they know how to regenerate themselves," says Hoegh-Guldberg. "Corals are almost made for this."

Amber Dance is a freelance writer in Los Angeles, California.

1. National Academies of Sciences, Engineering, and Medicine. *A Research Review of Interventions to Increase the Persistence and Resilience of Coral Reefs* (2019).
2. Page, C. A., Muller, E. M. & Vaughan, D. E. *Ecol. Eng.* **123**, 86–94 (2018).
3. Darling, E. S. et al. *Nature Ecol. Evol.* **3**, 1341–1350 (2019).
4. Reid, E. C. et al. *Limnol. Oceanogr.* **64**, 1949–1965 (2019).
5. Beyer, H. L. et al. *Conserv. Lett.* **11**, e12587 (2018).
6. Pörtner, H.-O. et al. *Special Report on the Ocean and Cryosphere in a Changing Climate* (IPCC, 2019).



Minecraft's open-ended play environment could be ideal for AI research, some researchers say.

AI VERSUS MINECRAFT

A coding contest aims to spur progress in machine-learning techniques. **By Jeremy Hsu**

To see the divide between the best artificial intelligence and the mental capabilities of a seven-year-old child, look no further than the popular video game *Minecraft*. A young human can learn how to find a rare diamond in the game after watching a 10-minute demonstration on YouTube. Artificial intelligence (AI) is nowhere close. But in a unique computing competition ending this month, researchers hope to shrink the gap between machine and child – and in doing so, help to reduce the computing power needed to train AIs.

Competitors may take up to four days and

use no more than eight million steps to train their AIs to find a diamond. That's still a lot longer than it would take a child to learn, but much faster than typical AI models nowadays.

The contest is designed to spur advances in an approach called imitation learning. This contrasts with a popular technique known as reinforcement learning, in which programs try thousands or millions of random actions in a trial-and-error fashion to home in on the best process. Reinforcement learning has helped generate recommendations for Netflix users, created ways to train robotic arms in factories and even bested humans in gaming. But it can require a lot of time and computing power.

Attempts to use reinforcement learning to create algorithms that can safely drive a car or win sophisticated games such as Go have involved hundreds or thousands of computers working in parallel to collectively run hundreds of years' worth of simulations – something only the most deep-pocketed governments and corporations can afford.

Imitation learning can improve the efficiency of the learning process, by mimicking how humans or even other AI algorithms tackle the task. And the coding event, known as the MineRL (pronounced 'mineral') Competition, encourages contestants to use this technique to teach AI to play the game.

Reinforcement-learning techniques wouldn't stand a chance in this competition on their own, says William Guss, a PhD candidate in deep-learning theory at Carnegie Mellon University in Pittsburgh, Pennsylvania, and head of the MineRL Competition's organizing team. Working at random, an AI might succeed only in chopping down a tree or two in the eight-million-step limit of the competition – and that is just one of the prerequisites for creating an iron pickaxe to mine diamonds in the game. "Exploration is really, really difficult," Guss says. "Imitation learning gives you a good prior about your environment."

Guss and his colleagues hope that the contest, which is sponsored by Carnegie Mellon and Microsoft among others, could have an impact beyond locating *Minecraft* gems, by inspiring coders to push the limits of imitation learning. Such research could ultimately

Feature

help to train AI so that it can interact better with humans in a wide range of situations, as well as navigate environments that are filled with uncertainty and complexity. “Imitation learning is at the very core of learning and the development of intelligence,” says Oriol Vinyals, a research scientist at Google DeepMind in London and a member of the MineRL Competition advisory committee. “It allows us to quickly learn a task without the need to figure out the solution that evolution found ‘from scratch.’”

Gaming by example

The group behind the competition says that *Minecraft* is particularly good as a virtual training ground. Players of the game showcase many intelligent behaviours. In its popular survival mode, they must defend themselves against monsters, forage or farm food and continually gather materials to build structures and craft tools. New players must learn *Minecraft*'s version of physics, as well as discover recipes to transform materials into resources or tools. The game has become famous for the creativity it unleashes in its players, who construct blocky virtual versions of a wide variety of things: the Eiffel Tower, Disneyland, the Death Star trench run from Star Wars, and even a working computer inside the game.

To create training data for the competition, MineRL organizers set up a public *Minecraft* server and recruited people to complete challenges designed to demonstrate specific tasks, such as crafting various tools. They ultimately captured 60 million examples of actions that could be taken in a given situation and approximately 1,000 hours of recorded behaviour to give to the teams. The recordings represent one of the first and largest data sets devoted specifically to imitation-learning research.

The contest focuses on using imitation to ‘bootstrap’ learning, so that AIs don’t need to spend so much time exploring the environment to find out what is possible from first principles, and instead use the knowledge that humans have built up, says Rohin Shah, a PhD candidate in computer science at the University of California, Berkeley, who runs the AI-focused *Alignment Newsletter*. “To my knowledge, there hasn’t been another AI competition focused on this question in particular.”

Spurred by cloud computing and an ample supply of data, reinforcement learning has typically generated the lion’s share of new AI research papers. But interest in imitation learning is picking up, in part because researchers are grappling with the limits of the trial-and-error approach. Learning in that way requires training data that can showcase all possibilities and consequences of different environmental interactions, says Katja Hofmann, principal researcher at the Game Intelligence group at Microsoft Research in

Cambridge, UK, and a member of the MineRL Competition’s organizing committee (Microsoft acquired *Minecraft*'s developer for US\$2.5 billion in 2014). Such data can be hard to come by in complex, real-world environments, in which it’s not easy or safe to play out all the consequences of bad decisions.

Take self-driving cars, for example. Training them mainly through reinforcement learning would require thousands or millions of trials to work out the differences between safe and reckless driving. But driving simulations cannot include all the possible conditions that could lead to a crash in the real world. And allowing a self-driving car to learn by crashing

“I’m particularly interested in *Minecraft* because it’s an example of an environment in which humans actually have diverse goals.”

repeatedly on public roads would be downright dangerous. Beyond the safety issues, reinforcement learning can get expensive, demanding computing power worth millions of dollars, Hofmann says.

Unlike pure reinforcement learning’s from-scratch approach, imitation learning takes short cuts, getting a head start by learning from example. It has already found a home in uses alongside reinforcement learning. Some of the most celebrated AI demonstrations of the past few years, including the AlphaGo algorithm’s 2017 trouncing of human Go masters, combined the two approaches, starting with a foundational model generated using imitation learning.

Imitation learning has limitations, too. One is that it is biased toward solutions that have already been demonstrated in the learning examples. AI trained in this way can therefore be inflexible. “If the AI system makes a mistake, or diverges somewhat from what a human would do, then it ends up in a new setting that’s different from what it saw in the demonstrations,” Shah says. “Since it hasn’t seen this situation, it becomes even more confused, and makes more mistakes, which compound further, leading to pretty bad failures.”

Still, a number of researchers see great potential in the technique, especially when it comes to training an AI to pursue specific objectives. “The nice part about imitation learning as opposed to reinforcement learning is you get demonstrations of success,” says Debadeepta Dey, principal researcher in the Adaptive Systems and Interaction group at Microsoft Research in Redmond, Washington. “This really helps to speed up learning.”

To get to the diamond treasure, the AI-controlled players, or agents, in the MineRL contest have to master a multi-step process.

First, they collect wood and iron to make pickaxes. Then they build torches to light the way. They might also carry a bucket of water to quench underground lava flows. Once all that is prepared, an AI can begin exploring mining shafts and caves, as well as tunnelling its way underground to search for diamond ore.

Competitors must train their AIs with a set of hardware consisting of no more than six central-processing cores and one NVIDIA graphics card – something that most research labs can afford through cloud-computing services. More than 900 teams signed up for the competition’s first round and 39 ended up submitting AI agents. The ten groups that made the most progress in training AIs to discover diamonds have advanced to the second and final round. Some of those AIs have managed to obtain iron ore and construct a furnace, two other prerequisites for making an iron pickaxe. But Guss doesn’t expect any of the teams’ agents to find a diamond – at least in this first competition.

Although the contest targets a specific objective, it could spur wider AI research with *Minecraft*. “I’m particularly interested in *Minecraft* because it’s an example of an environment in which humans actually have diverse goals – there’s no ‘one thing’ that humans do in *Minecraft*,” Shah says. “This makes it a much more appropriate test bed for techniques that attempt to learn human goals.”

And even if the game’s graphics and rules do not perfectly reflect physical reality, developing more-efficient ways of training AIs in *Minecraft* could translate to speedier AI learning in areas such as robotics. MineRL “could lead to results which would have an impact in real-world domains, such as robotic assembly of complex objects or any other domain where learning complex behaviour is required”, says Joni Pajarinen, a research group leader in the Intelligent Autonomous Systems lab at the Technical University of Darmstadt in Germany.

Once the final round of the competition wraps up on 25 November, Guss and other organizers will review the submissions to determine which AI proves the most advanced diamond hunter. The final results will become public on 6 December, just before NeurIPS (the Conference on Neural Information Processing Systems) in Vancouver, Canada, where all ten finalist teams are invited to present their results.

If the MineRL Competition catches on and becomes a recurring tradition, it could provide a public benchmark for tracking progress in imitation learning. “It seems quite likely that MineRL will encourage more research into imitation learning,” Shah says. “Whether imitation learning will have significance for real-world applications remains to be seen, but I am optimistic.”

Jeremy Hsu is a freelance journalist based in New York City.

Books & arts



In the film *Terminator: Dark Fate*, Linda Hamilton plays Sarah Connor as an older woman — a demographic that's rarer in science-fiction novels.

Space ageing: where are the galactic grandmas?

The lack of older women in sci-fi novels reflects and reifies ageism and sexism. **By Sylvia Spruck Wrigley**

As women get old, they gain a superpower: invisibility. And not only in real life. 'Young adult' fantasy and science-fiction hits such as Suzanne Collins's novel series *The Hunger Games* and Stephenie Meyers's *Twilight* series have been taken to task for doing away with

mature women. In fantasy generally, older women mainly occupy supporting roles, such as fairy godmothers, wise crones and evil witches. The best are subversions — George R. R. Martin's *Queen of Thorns* in *A Song of Ice and Fire*, for instance, or Terry Pratchett's wonderful *Granny Weatherwax* and *Nanny Ogg* in

the *Discworld* series. All of them embrace old age with gusto.

I expected better from science-fiction novels, where alternative worlds and alien nations explore what it means to be human. In 1976, after all, Ursula K. Le Guin argued in her essay 'The Space Crone' that post-menopausal women are best suited to representing the human race to alien species, because they are the most likely to have experienced all the changes of the human condition. And Robert A. Heinlein offers a fantastic galactic grandmother in *The Rolling Stones* (1952): Hazel Stone, engineer, lunar colonist and expert blackjack player irritated by the everyday misogyny of the Solar System.

Over the past year, with support from such authors and readers all over the world, I've searched for competent, witty female elders in major roles in sci-fi novels. I found no shortage



In Stephen King's *The Stand*, Abigail Freemantle becomes a spiritual leader as a centenarian.

of fantastic female characters across the genre, from the gynocentric utopians of Charlotte Perkins Gilman's 1915 *Herland* to mathematician Elma York in Mary Robinette Kowal's 2018 *The Calculating Stars*. But I have so far confirmed just 36 English-language novels in the genre that feature old women as major figures. The earliest is Gertrude Atherton's 1923 *Black Oxen*; the latest, from 2018, are *Blackfish City* by Sam J. Miller and *Record of a Spaceborn Few* by Becky Chambers.

Experience gap

It's a notable gap. Old men with deep expertise and experience throng sci-fi, from ancient keeper to eccentric mentor, retired badass and wasteland elder. And non-binary gender possibilities are explored in books such as Octavia

Butler's 1987 *Dawn* and Kameron Hurley's *The Mirror Empire* (2014).

Over the past century, women in the real world have been increasingly likely to become researchers, doctors and engineers. Indeed, most fields in science, technology, engineering and mathematics now recruit a growing proportion of women. But women are still under-represented among senior scientists, owing to the 'leaky pipeline' – they leave the field disproportionately in response to systemic bias.

And science fiction has magnified that issue. Rather than countering bias against ageing women, sci-fi writers seem more interested in making them young again – even expediting the rejuvenation process by casting it as a modern convenience akin to jet packs and

replicators. Yet again, only a few of the novels I found featuring technological fountains of youth include old women. Paula Myo in Peter Hamilton's *Commonwealth Saga* (2004) and Sarah Halifax in Robert Sawyer's *Rollback* (2007), for example, consider the side effects of gaining life experience without apparent ageing. John Scalzi's *Old Man's War* (2005) is a thought-provoking parody of the 'body-snatching' trope, in which a new body is taken to replace a worn one, showing the psychological perils of renewal that's only skin-deep.

“Rather than countering bias against ageing women, sci-fi writers seem more interested in making them young again.”

Age-reversal technology should apply to all genders: why would anyone get physically old if they didn't have to? And yet, amid the fictional horde of seasoned male mentors and stewards, sci-fi authors struggle to imagine a similar function for aged women.

All this reflects a general societal reluctance to see ageing as a natural process. The global anti-ageing market is currently worth more than US\$50 billion, mainly targeting women aged 35 to 55 in a kind of heckling by advertisement. Many women are reluctant to describe themselves as elderly from fear of stereotypes that define older women as isolated and fragile. If a woman is smart and social and competent, the stereotypes say, she must not be old.

Double dearth

I found a clear lack of cultural diversity in the female elders of sci-fi, almost as if there's a quota. Even when I focused on Afrofuturism, searching the works of Butler and fellow pioneer Samuel R. Delany, I found older women of colour only in short stories and fantasy novels. Many point to Le Guin's anarchist leader Laia Asieo Odo. However, Odo is middle-aged when described in *The Dispossessed* (1974), and old (and nearing death) only in the short story 'The Day Before the Revolution'. Similarly, Essun of N. K. Jemisin's *Broken Earth* trilogy is in her forties. I discovered just one major character who met all the criteria: Mother Abigail from Stephen King's *The Stand* (1978).

UK sci-fi authors, I found, consistently portray only white old women as British. Women of colour are always described as coming from other countries or realms; I have noted no representation of actual UK diversity. US writers have a similar record, with one exception. Three female elders from Native nations feature prominently in novels:

Masaaraq from *Blackfish City*; Jenny Casey from Elizabeth Bear's *Hammered* (2004); and Kris Longknife in the eponymous series by Mike Shepherd.

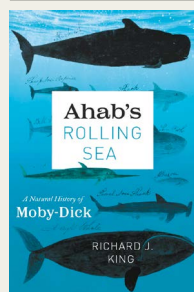
Celibacy is another distinguishing feature of the few female elders there are in sci-fi — despite studies showing that around half of women over 40 (including a significant number of over-80s) are sexually active and satisfied (S. E. Trompeter *et al. Am. J. Med.* **125**, 37–43.e1; 2012). Madame Zattiany of *Black Oxen*, for instance, remains uninterested in sex despite a glandular 'rejuvenation' that leads to an affair with a much younger man. (Despite the lack of libido, it is her inability to bear children that dooms the relationship.) Interestingly, the elderly female characters who show any interest in sex are clearly defined as lesbian or bisexual, such as the killer-whale-riding warrior grandmother in *Blackfish City*. As for menopause, it is glossed over in the qualifying books, in sharp contrast to the prevalence of puberty-related tales. That is a definite reflection of the dearth of research around menopause, and of support for women undergoing it.

I have shared my data through an open mailing list, and asked for input at presentations at major sci-fi conferences in Europe. Recently, I received an e-mail asking why I expected 'cronies' to appear in sci-fi at all. Shouldn't youth take centre stage, the author asked, with the added advantage of romantic potential? This summarizes the attitude that, after a certain age, women are uninteresting or threatening — and need to be got out of the way. More than 20% of US citizens will be 65 or over by 2035. The real-world grey tsunami cannot be halted. It could be considered a blessing, if we were to collectively focus on the strengths of older people, and foster healthspan as well as lifespan.

The wise old crone might not be a useful trope. However, science fiction could and should explore new roles for female elders as multifaceted beings. Authors have an opportunity here. With so few venerable women in major roles, a single novel including a new manifestation (say, a crime-solving octogenarian tribble rancher or a trans woman over 50) could completely change the landscape. Grandma is marvellous already; she doesn't need to look like a teenager.

Sylvia Spruck Wrigley is a speculative-fiction author and independent scholar based in Tallinn, who was nominated for a Nebula Award in 2014. Her short stories have been translated into more than a dozen languages. She is co-author of *The Triangle*, an audiobook production for the app Serial Box, set in the Bermuda Triangle. Find out more about her work on Old Women in Science Fiction by joining the mailing list at <https://intrigue.co.uk>. e-mail: sylvia@intrigue.co.uk

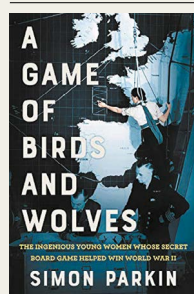
Books in brief



Ahab's Rolling Sea

Richard J. King Univ. Chicago Press (2019)

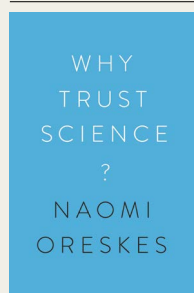
Herman Melville's sprawling masterpiece *Moby-Dick* (1851) is a fictional feat studded with empirical evidence, reveals maritime historian Richard King in this invigorating study. King traces references to ethology, meteorology, marine microbiota and the oceans to Melville's sailing experience in the Pacific and wranglings with the works of scientists William Scoresby, Louis Agassiz and others. *Moby-Dick*, King boldly avers, is a "proto-Darwinian fable" — and its beleaguered narrator, Ishmael, an early environmentalist.



A Game of Birds and Wolves

Simon Parkin Sceptre (2019)

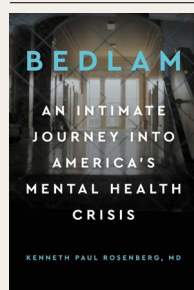
Did gaming win the Second World War? In this stirring history, Simon Parkin recounts how eight mathematically minded members of the UK Women's Royal Naval Service, with retired captain Gilbert Roberts, aimed to crack the tactics of Germany's notorious U-boats through war games. Playing large-scale Battleship on the floor of a Liverpool office, the team's 'Operation Raspberry' was decisive in winning the Battle of the Atlantic. Parkin's account redresses a balance: none in this doughty sisterhood has ever been publicly honoured.



Why Trust Science?

Naomi Oreskes Princeton Univ. Press (2019)

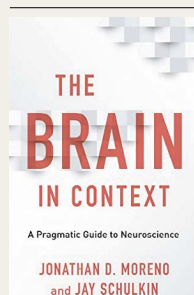
As some sectors of society reject expertise on issues such as vaccination, science historian Naomi Oreskes explores what makes science trustworthy. This concise volume — comprising her 2016 Tanner Lectures on Human Values at Princeton University in New Jersey, along with commentary by experts — is a bracing exploration of philosophy of science and a demonstration of her vigorous engagement with the topic. We trust science, she reminds us, because consensus is a crucial indicator of truth — and "objectivity is maximised" through diversity.



Bedlam

Kenneth Paul Rosenberg Avery (2019)

Psychiatrist Kenneth Rosenberg has been at the front lines of mental illness since the 1980s, when US psychiatric-hospital closures forced many people with serious mental conditions onto the streets or into prisons: some jails now 'warehouse' thousands. He meshes research with an analysis of systemic failures and personal stories, including those of psychiatrist Elyn Saks and his own sister, both diagnosed with schizophrenia. His ultimately hopeful study highlights key steps for patients, from details on integrated care to US legal advice.



The Brain in Context

Jonathan D. Moreno and Jay Schulkin Columbia Univ. Press (2019)

That fatty mass in the skull is not all there is to the brain — neural tissue lurks all over the body. So bioethicist Jonathan Moreno and neuroscientist Jay Schulkin begin their guide to neurology. To "see the brain in its wholeness", they examine the historical interplay of experiment and theory through lenses from comparative structure to evolution and imaging. The result is fascinating, whether on 'brains in a dish' or BrainGate technology to help people with paralysis control their limbs; but factual logjams impede the flow at times. **Barbara Kiser**

Masaaraq from *Blackfish City*; Jenny Casey from Elizabeth Bear's *Hammered* (2004); and Kris Longknife in the eponymous series by Mike Shepherd.

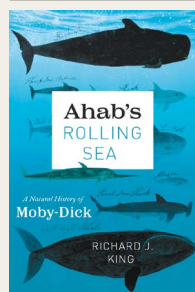
Celibacy is another distinguishing feature of the few female elders there are in sci-fi — despite studies showing that around half of women over 40 (including a significant number of over-80s) are sexually active and satisfied (S. E. Trompeter *et al. Am. J. Med.* **125**, 37–43.e1; 2012). Madame Zattiany of *Black Oxen*, for instance, remains uninterested in sex despite a glandular 'rejuvenation' that leads to an affair with a much younger man. (Despite the lack of libido, it is her inability to bear children that dooms the relationship.) Interestingly, the elderly female characters who show any interest in sex are clearly defined as lesbian or bisexual, such as the killer-whale-riding warrior grandmother in *Blackfish City*. As for menopause, it is glossed over in the qualifying books, in sharp contrast to the prevalence of puberty-related tales. That is a definite reflection of the dearth of research around menopause, and of support for women undergoing it.

I have shared my data through an open mailing list, and asked for input at presentations at major sci-fi conferences in Europe. Recently, I received an e-mail asking why I expected 'cronies' to appear in sci-fi at all. Shouldn't youth take centre stage, the author asked, with the added advantage of romantic potential? This summarizes the attitude that, after a certain age, women are uninteresting or threatening — and need to be got out of the way. More than 20% of US citizens will be 65 or over by 2035. The real-world grey tsunami cannot be halted. It could be considered a blessing, if we were to collectively focus on the strengths of older people, and foster healthspan as well as lifespan.

The wise old crone might not be a useful trope. However, science fiction could and should explore new roles for female elders as multifaceted beings. Authors have an opportunity here. With so few venerable women in major roles, a single novel including a new manifestation (say, a crime-solving octogenarian tribble rancher or a trans woman over 50) could completely change the landscape. Grandma is marvellous already; she doesn't need to look like a teenager.

Sylvia Spruck Wrigley is a speculative-fiction author and independent scholar based in Tallinn, who was nominated for a Nebula Award in 2014. Her short stories have been translated into more than a dozen languages. She is co-author of *The Triangle*, an audiobook production for the app Serial Box, set in the Bermuda Triangle. Find out more about her work on Old Women in Science Fiction by joining the mailing list at <https://intrigue.co.uk>. e-mail: sylvia@intrigue.co.uk

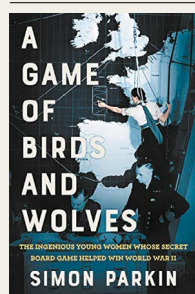
Books in brief



Ahab's Rolling Sea

Richard J. King Univ. Chicago Press (2019)

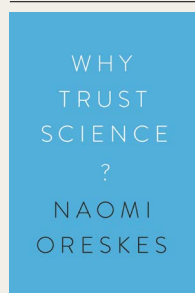
Herman Melville's sprawling masterpiece *Moby-Dick* (1851) is a fictional feat studded with empirical evidence, reveals maritime historian Richard King in this invigorating study. King traces references to ethology, meteorology, marine microbiota and the oceans to Melville's sailing experience in the Pacific and wranglings with the works of scientists William Scoresby, Louis Agassiz and others. *Moby-Dick*, King boldly avers, is a "proto-Darwinian fable" — and its beleaguered narrator, Ishmael, an early environmentalist.



A Game of Birds and Wolves

Simon Parkin Sceptre (2019)

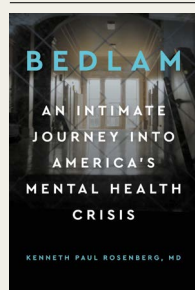
Did gaming win the Second World War? In this stirring history, Simon Parkin recounts how eight mathematically minded members of the UK Women's Royal Naval Service, with retired captain Gilbert Roberts, aimed to crack the tactics of Germany's notorious U-boats through war games. Playing large-scale Battleship on the floor of a Liverpool office, the team's 'Operation Raspberry' was decisive in winning the Battle of the Atlantic. Parkin's account redresses a balance: none in this doughty sisterhood has ever been publicly honoured.



Why Trust Science?

Naomi Oreskes Princeton Univ. Press (2019)

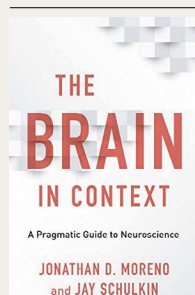
As some sectors of society reject expertise on issues such as vaccination, science historian Naomi Oreskes explores what makes science trustworthy. This concise volume — comprising her 2016 Tanner Lectures on Human Values at Princeton University in New Jersey, along with commentary by experts — is a bracing exploration of philosophy of science and a demonstration of her vigorous engagement with the topic. We trust science, she reminds us, because consensus is a crucial indicator of truth — and "objectivity is maximised" through diversity.



Bedlam

Kenneth Paul Rosenberg Avery (2019)

Psychiatrist Kenneth Rosenberg has been at the front lines of mental illness since the 1980s, when US psychiatric-hospital closures forced many people with serious mental conditions onto the streets or into prisons: some jails now 'warehouse' thousands. He meshes research with an analysis of systemic failures and personal stories, including those of psychiatrist Elyn Saks and his own sister, both diagnosed with schizophrenia. His ultimately hopeful study highlights key steps for patients, from details on integrated care to US legal advice.



The Brain in Context

Jonathan D. Moreno and Jay Schulkin Columbia Univ. Press (2019)

That fatty mass in the skull is not all there is to the brain — neural tissue lurks all over the body. So bioethicist Jonathan Moreno and neuroscientist Jay Schulkin begin their guide to neurology. To "see the brain in its wholeness", they examine the historical interplay of experiment and theory through lenses from comparative structure to evolution and imaging. The result is fascinating, whether on 'brains in a dish' or BrainGate technology to help people with paralysis control their limbs; but factual logjams impede the flow at times. **Barbara Kiser**

Comment



QILAI SHEN/BLOOMBERG/GETTY

A researcher stacks mouse containers at an animal-breeding facility near Guangzhou, China.

Five ways China must cultivate research integrity

Li Tang

A swift increase in scientific productivity has outstripped the country's ability to promote rigour and curb academic misconduct; it is time to seize solutions.

How researchers in China behave has an impact on the global scientific community. With more than four million researchers, China has more science and technology personnel than any other nation. In 2008, it overtook the United Kingdom in the number of articles indexed in the Web of Science, and now ranks second in the world. In 2018, China published 412,000 papers.

But China also produces a disproportionate number of faked peer reviews and plagiarized or fraudulent publications. Its share of retracted papers is around three times that expected from its scientific output (see 'Outsized retractions').

The past few years have witnessed high-profile cases of faked peer reviews, image manipulations and authorships for sale, some involving prominent Chinese scientists. In May last year, China asked two groups to foster research integrity and manage misconduct cases: its Ministry of Science and Technology (MOST) and the Chinese Academy of Social Sciences (CASS). In November 2018, 41 national government agencies endorsed a set of 43 penalties for major academic misconduct. These range from terminating grants to restricting academic promotion and revoking business licences. This year, the government issued a foundational

document to promote the scientific enterprise and foster a culture of academic integrity¹.

China's strides towards reform have been well received domestically and abroad, but effecting lasting change is hard². To better characterize the situation, my team has studied global retraction data alongside national grants and applications that were revoked. We also surveyed researchers online and interviewed major stakeholders in China^{3,4}. These included experts on university ethics committees, programmes for research-integrity training and plagiarism detection, as well as funding-programme managers, journal editors and academics. Here, I outline major challenges in research integrity, and potential strategies and solutions to buttress it.

Five strategies

Align norms. What counts as misconduct rather than acceptable practice differs across cultural settings and disciplines. The lack of consensus over what misconduct means is a thorny challenge for an emerging scientific powerhouse. One of our interviewees noted that senior academics even disagreed over what constitutes an allegation.

Any discussion about misconduct and penalties is buffeted by conflicting norms: historical versus the present, national versus

Comment

international. For example, the reuse of text without proper citation is, to some degree, accepted in textbook publishing in China. Until 1999, duplicate submissions or even dual publication in Chinese and English were not considered particularly inappropriate. More than 20% of our survey respondents felt that duplicate submission and self-plagiarism were common in their domain. These are deemed misconduct in international scientific communities.

That presents Chinese scientific leaders with a dilemma: if wrongdoing is not punished, the scientific community could become more tolerant, and there might be more misconduct and recidivism. That would waste public money, erode trust in science and tarnish the country's reputation. Already, Chinese academics can find it difficult to maintain or expand international collaborations, and universities and funding agencies outside China have ethical concerns about forming partnerships.

But requiring strict compliance with international norms would target a broad spectrum of misbehaviours that are common practice. And high standards with unworkable rules could legitimize non-compliance⁵. Either scenario could stymie reform.

Optimize approaches. Research misbehaviour needs to be policed. Strategies can be classed as 'patrols' or 'fire alarms'⁶. Like other countries, China deploys both.

On the patrol side, China National Knowledge Infrastructure (CNKI), a Chinese version of the Web of Science database, provides a plagiarism-checking service to Chinese journals and universities. These have deployed CNKI software to detect plagiarized texts, including those saved as manipulated images. Since 2010, grant proposals have been checked for possible plagiarism at the National Natural Science Foundation of China (NSFC). Similarly, the National Social Science Fund of China (NSSFC) instigates systematic clean-ups for its funded projects, halting those that are left unfinished after the completion deadline (typically six years after receiving the grant). This put an end to 302 of 5,035 grants funded from 2002 to 2005. Terminated projects increased from 60 in 2002 to 99 in 2005, but have plummeted since checks were implemented and publicized in 2012 (ref. 3; see 'Checks changed behaviour').

Patrol deters certain types of misconduct, particularly before a grant or degree is awarded or a paper accepted. But patrols require dedicated software and infrastructure, so are costly to enforce. Every May (just before graduation), college students, university faculty members and support staff spend hours checking theses for plagiarism.

Perhaps that is why a fire-alarm tactic is dominant. China's science agencies and universities often wait to act until contacted by the media, wronged parties or whistle-blowers, and they focus most on cases that grab headlines. This

can be effective in the short term: in 2017, after 107 articles by Chinese authors were retracted by the journal *Tumor Biology* for faked peer reviews, investigations were completed within 4 months. More than 100 people were penalized and some 40 NSSFC grants revoked. But the fire-alarm tactic leads to selective investigations and uncertainty. It punishes past offences, but does little to deter future ones.

Empower enforcement. The burden of policing misconduct is too much for national agencies in any country, China included. That power is delegated to the universities and institutes where researchers work. But these organizations, concerned about soiling reputations and losing grant funds, are often unwilling to investigate alleged misconduct. They tend to respond only when whistles are blown. That depends on whistle-blowers who shoulder great professional and personal risk, especially in Chinese society, which values collectivism and interdependence over individualism and independence. In a 2017 survey of Chinese scholars, more than half of respondents who observed misconduct in the past three years said that they did nothing about it (unpublished results; see also Supplementary Information).

Assign responsibility. Perhaps the most difficult challenge in China, as elsewhere, is whether and to what extent to hold team members accountable for misconduct in joint work. Increasing specialization and globalization has made collaborations larger and more essential. That complicates how to allocate blame as well as credit. Should each listed author be held accountable for the full work, or just for their own? Should the corresponding author take most of the responsibility for fraud and errors others committed? Although more journals are requiring detailed descriptions of authors' contributions, discerning who should be responsible for a collaborative piece of work is difficult. This is particularly true when older articles are

retracted as a result of proven fraud – often, author contributions have not been specified.

The supervisor–student relationship poses a particular dilemma. In China, when PhD students are found guilty of misconduct, their supervisors are also punished. In recent scandals, plagiarists were stripped of their doctoral degrees, and their supervisors were demoted and barred from taking on PhD candidates. Alternatively, junior scientists might be punished, while senior ones responsible for misconduct retain status and position. Some argue that holding members of a research team accountable by association will improve enforcement and prevent scapegoating; others say that this shift in responsibility is unfair and burdensome.

Cultivate integrity. China's rapid research development must be brought into sync with a culture of integrity. Like other countries, it has seen that tying publication requirements to degree requirements, promotion or monetary rewards can lure researchers into inappropriate behaviour⁷.

Integrated tactics

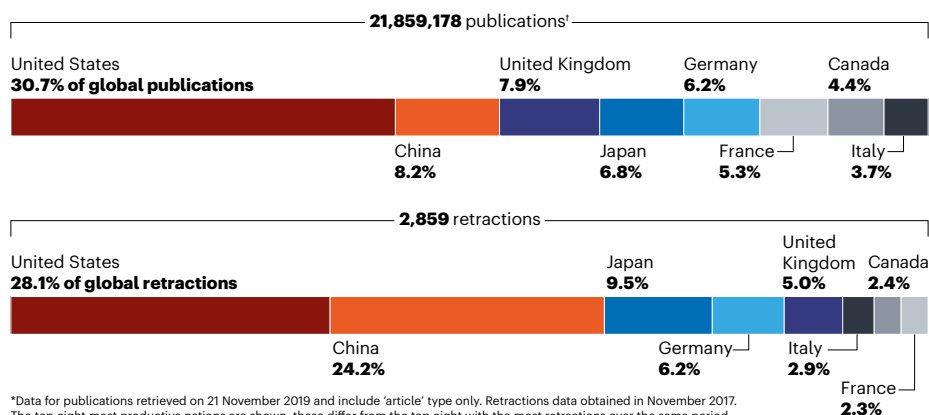
What is the best way to implement these strategies? I propose that working on several fronts will make each easier to accomplish.

Forgive, then be tough. China's scientific community first needs to agree on a common code of academic integrity that defines misconduct and undesirable research practices and sets out sanctions. China has a greater diversity of funders and a more mobile scientific workforce than ever before, so all stakeholders – including researchers, managers, journal editors and funding officers – must be in accord.

Penalties should focus on the most egregious acts, which are universally recognized: falsification, fabrication, plagiarism and fake reviews. Researchers should be admonished for past fraud but face harsher penalties for incidents that occur once the code is in place.

OUTSIZED RETRACTIONS

China has published 8% of the world's scientific articles, but by 2017 had garnered 24% of all retractions*.



SOURCE: L. TANG/WEB OF SCIENCE

Less serious questionable practices that were historically accepted should be subject to a statute of limitations.

Institutionalize. Integrity must be built into scientific institutions, with MOST and CASS taking the lead. CASS should set up departments to oversee misconduct cases, as MOST has. Both agencies should facilitate communication between all stakeholders and coordinate input from research societies to formulate workable rules that are compatible with international norms.

Transparency will help. Funding agencies should, for example, publicize the claimed achievements and promised research outputs of award recipients in prestigious talent programmes. This accountability will deter fraud and false advertising. China's General Administration of Press and Publications can help by urging Chinese publishers and database providers to take a proactive stance. For instance, Chinese journals often simply remove retracted articles from their collections and the CNKI database. Instead, journals should explicitly mark articles as retracted, as many Western journals do⁸. They should also share their 'blacklists' of authors who have repeatedly been found guilty of duplicate submissions.

With the right support, universities and research institutes can be best placed to initiate misconduct investigations. MOST and CASS should help them set up procedures. These should include appointing an independent ombudsperson to protect whistle-blowers and those accused of misconduct, for example by developing strategies to prevent cyberbullying and smear campaigns. In addition, each university should employ a professional chief integrity officer – not a faculty 'volunteer' – who reports directly to a vice-president.

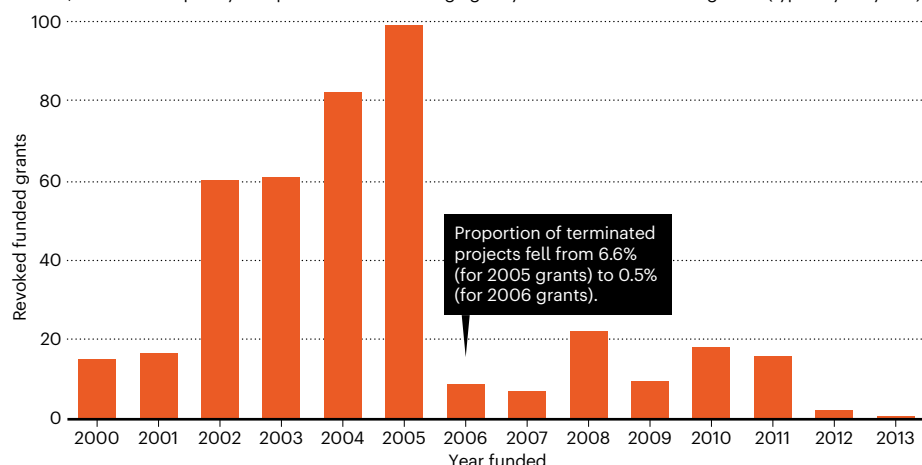
Incentivize. Administrative agencies must explicitly link support for a university to whether it vigorously investigates misconduct allegations and promotes integrity education, including putting dedicated professionals in place. Agencies can also set up open, regular communication about reform with junior and senior researchers – for real-world input and to allow institutions to learn from each other.

Educate. A healthy academic atmosphere cannot be built on penalties for misbehaviour alone. Universities could set up research-integrity help desks and hotlines, making contact information and investigation procedures accessible. The Chinese university code of academic integrity should be linked from every course syllabus. Teachers should have access to plagiarism-checking software and to training so they can understand its shortcomings.

More broadly, universities must work out how to provide effective integrity education. Training upstream is always better than

CHECKS CHANGED BEHAVIOUR

After the National Social Science Fund of China began terminating grants for incomplete overdue projects in 2012, researchers quickly complied with the funding agency's deadlines for finishing work (typically six years).



SOURCE: NSFC

disciplining transgressors after the fact (see also go.nature.com/2rpdhkv).

Many Chinese universities now require graduate students to take responsible-conduct courses. Around three-quarters of our survey respondents said they had received training in research ethics and integrity. Those enrolling for a PhD at Fudan University in Shanghai, for example, must attend mandatory ethics modules. Only those who pass the ethics quiz can register for further coursework.

Such training needs to be universal across Chinese institutions, and at all levels: for faculty members, technicians and non-scientific staff. Principal investigators who coordinate collaborations, as well as young researchers who collect, check and validate data, must know and accept their responsibilities⁴. 'Trust and verify' should be bywords for all. For example, at least two team members should collect and code raw data and record source links and detailed procedures. Pre-registration of analysis plans could also prevent tampering⁹.

Study. Also needed is rigorous research on what kind of institutional structures and programmes foster integrity, which types of training effect the most lasting change, and how to apply best practice. Comparative studies could provide lessons from other countries that have experience in combating academic misconduct and cultivating integrity. For example, in 2014, Denmark adopted a new code of conduct for research integrity as a result of orchestrated efforts by researchers, funding agencies and other stakeholders. The Netherlands followed suit in 2018. Indian efforts against predatory publishing could be adapted for China, as could the long-established US emphasis on quality rather than quantity in research evaluation.

To gather this knowledge, oversight agencies should have an open-door policy for stakeholders to express constructive and diverse opinions. Proceedings of misconduct investigations should be made public, not be shrouded in

secrecy¹⁰. Funding agencies need to earmark money for research-integrity studies to attract bright minds to the field. This year, the NSFC issued an open call for proposals on research integrity and ethics; it is unclear whether such funds will be available in future.

China must curb misconduct and foster integrity if it is to realize the central government's ambition of "world-class universities, world-class disciplines". It is still too early to anticipate all the changes reforms will bring, but the government has signalled its determination to act. We might see more investigations of misconduct because of closer scrutiny in the next couple of years. Improving the research practices of Chinese scholars will boost innovation and development everywhere.

The author

Li Tang is a professor of public policy at the School of International Relations and Public Affairs, Fudan University, Shanghai, China; and a member of the board of directors of the Chinese Association for Science of Science and S&T Policy (CASSSP).
e-mail: litang@fudan.edu.cn

- General Offices of the Communist Party of China Central Committee and the State Council. [in Chinese] 'Opinions on Further Promoting the Spirit of Scientists and Strengthening the Style of Work and Study Style' (CPC Central Committee & State Council, 2019); available at <https://go.nature.com/2d29xhj>
- Lei, R., Zhai, X., Zhu, W. & Qiu, R. *Nature* **569**, 184–186 (2019).
- Tang, L. & Wang, L. [in Chinese] *Sci. Sci. Manag. S&T* **40**, 15–30 (2019).
- Walsh, J. P., Lee, Y.-N. & Tang, L. *Res. Pol.* **48**, 444–461 (2019).
- Pedro, A. C. *Rational. Soc.* **30**, 80–107 (2018).
- McCubbins, M. D. & Schwartz, T. *Am. J. Polit. Sci.* **28**, 165–179 (1984).
- Fanelli, D., Costas, R. & Larivière, V. *PLoS ONE* **10**, e0127556 (2015).
- Harrison, W. *et al. Acta Cryst.* **E66**, e1–e2 (2010).
- Nosek, B. A., Ebersole, C. A., DeHaven, A. C. & Mellor, D. T. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
- Gunsalus, C. *Nature* **570**, 7 (2019).

Supplementary Information accompanies this article: see go.nature.com/2qahhbu

Climate tipping points — too risky to bet against

Timothy M. Lenton, Johan Rockström, Owen Gaffney, Stefan Rahmstorf, Katherine Richardson, Will Steffen & Hans Joachim Schellnhuber

The growing threat of abrupt and irreversible climate changes must compel political and economic action on emissions.

Politicians, economists and even some natural scientists have tended to assume that tipping points¹ in the Earth system — such as the loss of the Amazon rainforest or the West Antarctic ice sheet — are of low probability and little understood. Yet evidence is mounting that these events could be more likely than was thought, have high impacts and are interconnected across different biophysical systems, potentially committing the world to long-term irreversible changes.

Here we summarize evidence on the threat of exceeding tipping points, identify knowledge gaps and suggest how these should be plugged. We explore the effects of such large-scale changes, how quickly they might unfold and whether we still have any control over them.

In our view, the consideration of tipping points helps to define that we are in a climate emergency and strengthens this year's chorus of calls for urgent climate action — from schoolchildren to scientists, cities and countries.

The Intergovernmental Panel on Climate Change (IPCC) introduced the idea of tipping points two decades ago. At that time, these 'large-scale discontinuities' in the climate system were considered likely only if global warming exceeded 5 °C above pre-industrial levels. Information summarized in the two most recent IPCC Special Reports (published in 2018 and in September this year)^{2,3} suggests that tipping points could be exceeded even between 1 and 2 °C of warming (see 'Too close for comfort').

If current national pledges to reduce greenhouse-gas emissions are implemented — and that's a big 'if' — they are likely to result in at least 3 °C of global warming. This is despite the goal of the 2015 Paris agreement to limit warming to well below 2 °C. Some economists,

assuming that climate tipping points are of very low probability (even if they would be catastrophic), have suggested that 3 °C warming is optimal from a cost–benefit perspective. However, if tipping points are looking more likely, then the 'optimal policy' recommendation of simple cost–benefit climate-economy models⁴ aligns with those of the recent IPCC report². In other words, warming must be limited to 1.5 °C. This requires an emergency response.

Ice collapse

We think that several cryosphere tipping points are dangerously close, but mitigating greenhouse-gas emissions could still slow down the inevitable accumulation of impacts and help us to adapt.

Research in the past decade has shown that the Amundsen Sea embayment of West Antarctica might have passed a tipping point³: the 'grounding line' where ice, ocean and bed-rock meet is retreating irreversibly. A model study shows³ that when this sector collapses, it could destabilize the rest of the West Antarctic ice sheet like toppling dominoes — leading to about 3 metres of sea-level rise on a timescale of centuries to millennia. Palaeo-evidence shows that such widespread collapse of the West Antarctic ice sheet has occurred repeatedly in the past.

The latest data show that part of the East Antarctic ice sheet — the Wilkes Basin — might be similarly unstable³. Modelling work suggests that it could add another 3–4 m to sea level on timescales beyond a century.

The Greenland ice sheet is melting at an accelerating rate³. It could add a further 7 m to sea level over thousands of years if it passes a particular threshold. Beyond that, as the elevation of the ice sheet lowers, it melts further, exposing the surface to ever-warmer air. Models suggest that the Greenland ice sheet could be doomed at 1.5 °C of warming³, which could happen as soon as 2030.

Thus, we might already have committed future generations to living with sea-level rises of around 10 m over thousands of years³. But that timescale is still under our control. The rate of melting depends on the magnitude of warming above the tipping point. At 1.5 °C, it could take 10,000 years to unfold³; above 2 °C it could take less than 1,000 years⁶.





An aeroplane flies over a glacier in the Wrangell St Elias National Park in Alaska.

Researchers need more observational data to establish whether ice sheets are reaching a tipping point, and require better models constrained by past and present data to resolve how soon and how fast the ice sheets could collapse.

Whatever those data show, action must be taken to slow sea-level rise. This will aid adaptation, including the eventual resettling of large, low-lying population centres.

A further key impetus to limit warming to 1.5 °C is that other tipping points could be triggered at low levels of global warming. The

“The clearest emergency would be if we were approaching a global cascade of tipping points.”

latest IPCC models projected a cluster of abrupt shifts⁷ between 1.5 °C and 2 °C, several of which involve sea ice. This ice is already shrinking rapidly in the Arctic, indicating that, at 2 °C of warming, the region has a 10–35% chance³ of becoming largely ice-free in summer.

Biosphere boundaries

Climate change and other human activities risk triggering biosphere tipping points across a range of ecosystems and scales (see ‘Raising the alarm’).

Ocean heatwaves have led to mass coral bleaching and to the loss of half of the shallow-water corals on Australia’s Great Barrier Reef. A staggering 99% of tropical corals are projected² to be lost if global average temperature rises by 2 °C, owing to interactions between warming, ocean acidification and pollution. This would represent a profound loss of marine biodiversity and human livelihoods.

As well as undermining our life-support system, biosphere tipping points can trigger abrupt carbon release back to the atmosphere. This can amplify climate change and reduce remaining emission budgets.

Deforestation and climate change are destabilizing the Amazon – the world’s largest rainforest, which is home to one in ten known species. Estimates of where an Amazon tipping point could lie range from 40% deforestation to just 20% forest-cover loss⁸. About 17% has been lost since 1970. The rate of deforestation varies with changes in policy. Finding the tipping point requires models that include deforestation and climate change as interacting drivers, and that incorporate fire and climate feedbacks as interacting tipping mechanisms across scales.

With the Arctic warming at least twice as quickly as the global average, the boreal forest in the subarctic is increasingly vulnerable. Already, warming has triggered large-scale insect disturbances and an increase



ALEXIS ROSENFELD/GETTY

Bleached corals on a reef near the island of Moorea in French Polynesia in the South Pacific.

in fires that have led to dieback of North American boreal forests, potentially turning some regions from a carbon sink to a carbon source⁹. Permafrost across the Arctic is beginning to irreversibly thaw and release carbon dioxide and methane – a greenhouse gas that is around 30 times more potent than CO₂ over a 100-year period.

Researchers need to improve their understanding of these observed changes in major ecosystems, as well as where future tipping points might lie. Existing carbon stores and potential releases of CO₂ and methane need better quantification.

The world's remaining emissions budget for a 50:50 chance of staying within 1.5 °C of warming is only about 500 gigatonnes (Gt) of CO₂. Permafrost emissions could take an estimated 20% (100 Gt CO₂) off this budget¹⁰, and that's without including methane from deep permafrost or undersea hydrates. If forests are close to tipping points, Amazon dieback could release another 90 Gt CO₂ and boreal forests a further 110 Gt CO₂ (ref. 11). With global total CO₂ emissions still at more than 40 Gt per year, the remaining budget could be all but erased already.

Global cascade

In our view, the clearest emergency would be if we were approaching a global cascade of tipping points that led to a new, less habitable, 'hothouse' climate state¹¹. Interactions

could happen through ocean and atmospheric circulation or through feedbacks that increase greenhouse-gas levels and global temperature. Alternatively, strong cloud feedbacks

could cause a global tipping point^{12,13}.

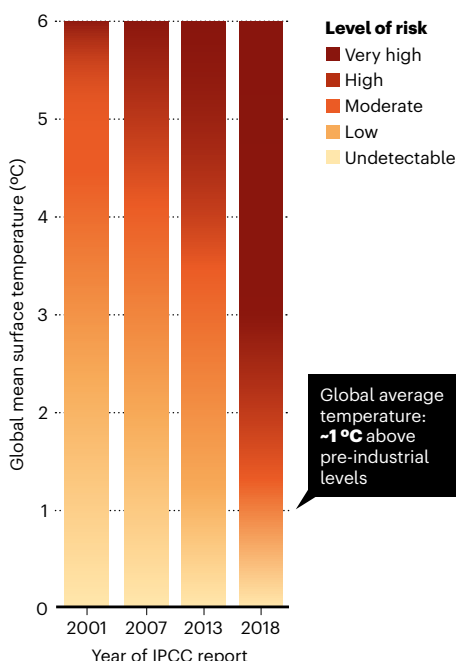
We argue that cascading effects might be common. Research last year¹⁴ analysed 30 types of regime shift spanning physical climate and ecological systems, from collapse of the West Antarctic ice sheet to a switch from rainforest to savanna. This indicated that exceeding tipping points in one system can increase the risk of crossing them in others. Such links were found for 45% of possible interactions¹⁴.

In our view, examples are starting to be observed. For example, Arctic sea-ice loss is amplifying regional warming, and Arctic warming and Greenland melting are driving an influx of fresh water into the North Atlantic. This could have contributed to a 15% slowdown¹⁵ since the mid-twentieth century of the Atlantic Meridional Overturning Circulation (AMOC), a key part of global heat and salt transport by the ocean³. Rapid melting of the Greenland ice sheet and further slowdown of the AMOC could destabilize the West African monsoon, triggering drought in Africa's Sahel region. A slowdown in the AMOC could also dry the Amazon, disrupt the East Asian monsoon and cause heat to build up in the Southern Ocean, which could accelerate Antarctic ice loss.

The palaeo-record shows global tipping, such as the entry into ice-age cycles 2.6 million years ago and their switch in amplitude and frequency around one million years ago,

TOO CLOSE FOR COMFORT

Abrupt and irreversible changes in the climate system have become a higher risk at lower global average temperatures.



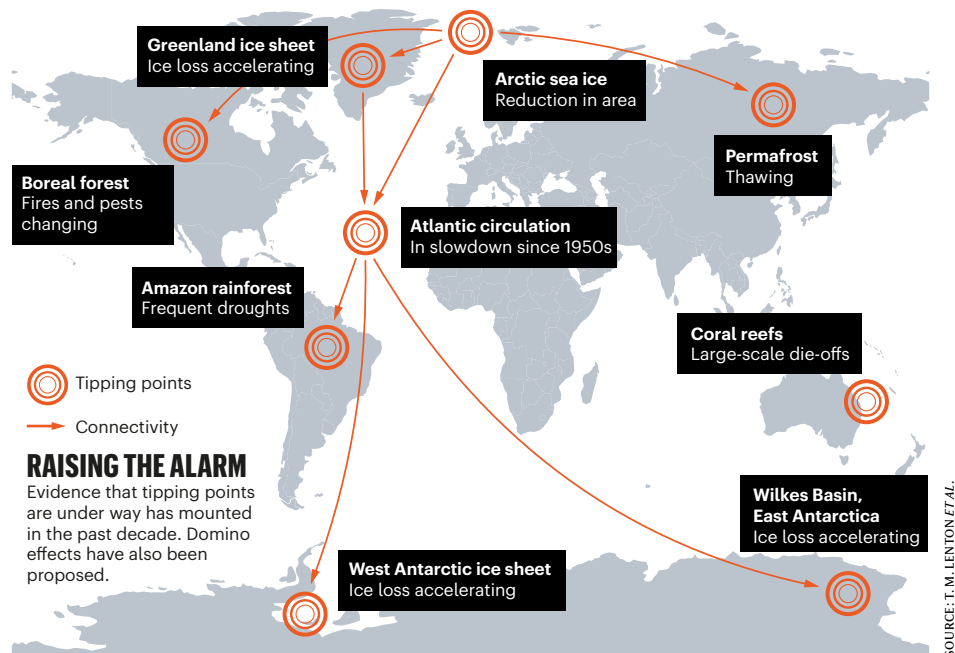
SOURCE: IPCC

which models are only just capable of simulating. Regional tipping occurred repeatedly within and at the end of the last ice age, between 80,000 and 10,000 years ago (the Dansgaard–Oeschger and Heinrich events). Although this is not directly applicable to the present interglacial period, it highlights that the Earth system has been unstable across multiple timescales before, under relatively weak forcing caused by changes in Earth's orbit. Now we are strongly forcing the system, with atmospheric CO₂ concentration and global temperature increasing at rates that are an order of magnitude higher than those during the most recent deglaciation.

Atmospheric CO₂ is already at levels last seen around four million years ago, in the Pliocene epoch. It is rapidly heading towards levels last seen some 50 million years ago – in the Eocene – when temperatures were up to 14 °C higher than they were in pre-industrial times. It is challenging for climate models to simulate such past 'hothouse' Earth states. One possible explanation is that the models have been missing a key tipping point: a cloud-resolving model published this year suggests that the abrupt break-up of stratocumulus cloud above about 1,200 parts per million of CO₂ could have resulted in roughly 8 °C of global warming¹².

Some early results from the latest climate models – run for the IPCC's sixth assessment report, due in 2021 – indicate a much larger climate sensitivity (defined as the temperature response to doubling of atmospheric CO₂) than in previous models. Many more results are pending and further investigation is required, but to us, these preliminary results hint that a global tipping point is possible.

To address these issues, we need models that capture a richer suite of couplings and feedbacks in the Earth system, and we need more data – present and past – and better ways to use them. Improving the ability of models to capture known past abrupt climate changes



and 'hothouse' climate states should increase confidence in their ability to forecast these.

Some scientists counter that the possibility of global tipping remains highly speculative. It is our position that, given its huge impact and irreversible nature, any serious risk assessment must consider the evidence, however limited our understanding might still be. To err on the side of danger is not a responsible option.

If damaging tipping cascades can occur and a global tipping point cannot be ruled out, then this is an existential threat to civilization. No amount of economic cost–benefit analysis is going to help us. We need to change our approach to the climate problem.

Act now

In our view, the evidence from tipping points alone suggests that we are in a state of planetary emergency: both the risk and urgency of the situation are acute (see 'Emergency: do the maths').

We argue that the intervention time left to prevent tipping could already have shrunk towards zero, whereas the reaction time to achieve net zero emissions is 30 years at best. Hence we might already have lost control of whether tipping happens. A saving grace is that the rate at which damage accumulates from tipping – and hence the risk posed – could still be under our control to some extent.

The stability and resilience of our planet is in peril. International action – not just words – must reflect this.

The authors

Timothy M. Lenton is director of the Global Systems Institute, University of Exeter, UK. **Johan Rockström** is director of

the Potsdam Institute for Climate Impact Research, Germany. **Owen Gaffney** is a global sustainability analyst at the Potsdam Institute for Climate Impact Research, Germany; and at the Stockholm Resilience Centre, Stockholm University, Sweden. **Stefan Rahmstorf** is professor of physics of the oceans at the University of Potsdam; and head of Earth system analysis at the Potsdam Institute for Climate Impact Research, Germany. **Katherine Richardson** is professor of biological oceanography at the Globe Institute, University of Copenhagen, Denmark. **Will Steffen** is emeritus professor of climate and Earth System science at the Australian National University, Canberra, Australia. **Hans Joachim Schellnhuber** is founding director of the Potsdam Institute for Climate Impact Research, Germany; and distinguished visiting professor, Tsinghua University, Beijing, China. e-mail: t.m.lenton@exeter.ac.uk

1. Lenton, T. M. et al. *Proc. Natl Acad. Sci. USA* **105**, 1786–1793 (2008).
2. IPCC. *Global Warming of 1.5°C* (IPCC, 2018).
3. IPCC. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* (IPCC, 2019).
4. Cai, Y., Lenton, T. M., & Lontzek, T. S. *Nature Clim. Change* **6**, 520–525 (2016).
5. Feldmann, J. & Levermann, A. *Proc. Natl Acad. Sci. USA* **112**, 14191–14196 (2015).
6. Aschwanden, A. et al. *Sci. Adv.* **5**, eaav9396 (2019).
7. Drijfhout, S. et al. *Proc. Natl Acad. Sci. USA* **112**, E5777–E5786 (2015).
8. Lovejoy, T. E. & Nobre, C. *Sci. Adv.* **4**, eaat2340 (2018).
9. Walker, X. J. et al. *Nature* **572**, 520–523 (2019).
10. Rogelj, J., Forster, P. M., Kriegler, E., Smith, C. J. & Séférian, R. *Nature* **571**, 335–342 (2019).
11. Steffen, W. et al. *Proc. Natl Acad. Sci. USA* **115**, 8252–8259 (2018).
12. Schneider, T., Kaul, C. M. & Pressel, K. G. *Nature Geosci.* **12**, 163–167 (2019).
13. Tan, I., Storelvmo, T. & Zelinka, M. D. *Science* **352**, 224–227 (2016).
14. Rocha, J. C., Peterson, G., Bodin, Ö. & Levin, S. *Science* **362**, 1379–1383 (2018).
15. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. *Nature* **556**, 191–196 (2018).

EMERGENCY: DO THE MATHS

We define emergency (E) as the product of risk and urgency. Risk (R) is defined by insurers as probability (p) multiplied by damage (D). Urgency (U) is defined in emergency situations as reaction time to an alert (τ) divided by the intervention time left to avoid a bad outcome (T). Thus:

$$E = R \times U = p \times D \times \tau / T$$

The situation is an emergency if both risk and urgency are high. If reaction time is longer than the intervention time left ($\tau/T > 1$), we have lost control.

Correspondence

Russia's stance on gene-edited humans

The Russian community of geneticists, clinicians and bioethicists have reached a consensus on the use of genome-editing technologies on human embryos and germ cells for clinical purposes (see, for example, *Nature* 574, 465–466; 2019). They consider that such experiments are premature at this point. Their view aligns with the position of the Russian ministry of health and sets the social context for further discussion of the technology.

We agree with the director-general of the World Health Organization (WHO) that comprehensive research is needed into the technical and ethical consequences of using the technology. We support the WHO advisory committee's recommendations to develop global standards for the governance and oversight of human-genome editing, and to create a public registry of clinical research on the effects of applying it to human somatic and germ cells (see *Nature* 575, 415–416; 2019).

Russian science recognizes the basic ethical principles that underpin the decisions of the United Nations, the United Nations Educational, Scientific and Cultural Organization, the WHO and other international organizations, as well as the provisions of the Council of Europe's Convention on Human Rights and Biomedicine. These principles will define the system of ethical expertise and inform how Russia regulates the field.

Elena G. Grebenshchikova

Russian Academy of Sciences,
Moscow, Russia.

aika45@yandex.ru

*On behalf of 14 co-signatories;
see go.nature.com/2jwmaq8.

We cannot all be ethicists

Sarah Franklin's message that "we must all be ethicists now" is laudable if it sensitizes researchers to the importance of ethical thinking (*Nature* 574, 627–630; 2019). We are concerned, however, that it could be misinterpreted to mean that expertise in ethics is no longer necessary for discussing issues pertaining to science and technology. This implication is dangerous in a society in which there is a mounting distrust of institutions and expert knowledge.

Franklin seems to us to conflate the field of enquiry of bioethics with bioethicists' participation in public bodies tasked with addressing science and technology governance. Expertise in bioethics cannot be improvised. Bioethicists are trained in the normative evaluation of biotechnologies, medical practices and other technologies. Bioethics aims to address questions related to what should or should not be done with regard to a particular issue. It is essential, therefore, that we protect the expertise that we have gathered through our training and experience.

Silvia Camporesi King's College
London, London, UK.
silvia.camporesi@kcl.ac.uk

Giulia Cavaliere Lancaster
University, UK.

Ukraine open index maps local citations

The Open Ukrainian Citation Index (OUCI; <http://ouci.dntb.gov.ua/en>) was launched this month by Ukraine's ministry of education and science, in conjunction with the country's State Scientific and Technical Library. This database could be particularly useful for tracking relationships between findings that concern regional topics and target domestic audiences, which are typically published in Ukrainian journals (see also A. J. Nederhof *Scientometrics* 66, 81–100; 2006).

Scholarly communication systems often fall short in revealing knowledge networks if their bibliographic and citation data are not in machine-readable form. The OUCI database, which comprises citations from all publishers that use Crossref's Cited-by service, corrects this problem. It upholds the aims of the Initiative for Open Citations, a collaboration of scholarly publishers, researchers and other stakeholders. It is accessible to researchers and the public, and it accounts for citations between publications (DOI to DOI) without the need to open the source articles. The database contains information from databases such as Scopus and the Web of Science, and can also be searched in English.

Dmytro Cheberkus Ministry of
Education and Science of Ukraine,
Kyiv, Ukraine.

Serhii Nazarovets State Scientific
and Technical Library of Ukraine,
Kyiv, Ukraine.
serhii.nazarovets@gmail.com

Jordan: networking across generations

Networks for young people interested in science rarely connect with societies for senior scientists, such as the American Association for the Advancement of Science, the Society for the Advancement of Science and Technology in the Arab World (SASTA) and The World Academy of Sciences (see A. Orben *Nature* 573, 465; 2019). As president of SASTA, I offer an example of a remedy from an unexpected source – Jordan, a country at the heart of crisis and displacement in the Middle East.

The Phi Science Institute is a network of young scientists across the Arab world. It holds an annual conference – Connect for Science – at which senior and junior scientists and students communicate with one another on an equal footing (<https://pris.phiscience.co>). The wisdom of the old meets the curiosity and enthusiasm of the young. Role models are set up for researchers at the start of their careers.

As mature scientists, we owe it to the next generation of researchers and to society to use our expertise to make a difference. I often invite well-known scientists to talk to my students, over Skype or in person. For example, US Nobel laureate Brian Kobilka shared his everyday experiences in the laboratory, and Magdalena Skipper, *Nature's* first female editor, told them her personal story.

Rana Dajani Hashemite University,
Zarqa, Jordan.
rdajani@hu.edu.jo

News & views

Cancer

Powerful system for cell protection revealed

Brent R. Stockwell

The discovery of a mechanism that guards against a type of cell death called ferroptosis reveals a system that regenerates a ubiquitous protective component of biological membranes, and might offer a target for anticancer drugs.

On 3 December 1956, the biochemist Frederick Crane detected, for the first time ever, a yellow substance purified from cow hearts, which had been obtained from the Oscar Mayer meat-processing factory in Madison, Wisconsin¹. In the laboratory of David Green at the University of Wisconsin, Crane investigated the oily material he had discovered, and struck gold: he found a lipid molecule that has a crucial role in energy generation in cells. But there were hints that it might have another function. Now, writing in *Nature*, Doll *et al.*² and Bersuker *et al.*³ report the discovery of this elusive role, finally revealing the missing part of the puzzle 63 years after Crane's discovery.

When Crane analysed the molecule he had identified, he noted that it was structurally similar to certain vitamins, and it was initially named vitamin Q₁₀. This molecule is important for respiration¹ – the energy-generating process that makes the molecule ATP – and acts in mitochondria, organelles that are responsible for much of the energy production in cells. It was subsequently discovered¹ to be ubiquitous. It not only resides in mitochondria, but is also present elsewhere in the cell in almost all lipid membranes.

The molecule was renamed ubiquinone because of its ubiquity and because it contains a type of chemical structure known as a quinone. Ubiquinone now also goes by the name coenzyme Q₁₀. But its function outside mitochondria has remained enigmatic, until now.

In 2012, a previously unknown type of programmed cell death called ferroptosis was described⁴. This iron-dependent pathway causes cells to die when a lipid modification called peroxidation degrades their membranes (Fig. 1). Ferroptosis has been linked to a range of processes. For example, it has been implicated

in degenerative diseases; in damage to plants caused by climate-change-associated heat stress; and in natural functions of the immune system, such as the elimination of tumour cells by killer T cells^{5–7}. Therapeutic approaches that harness ferroptosis are being developed for some cancers, given that the abnormalities of certain tumour cells can make them vulnerable to this type of cell death.

The damage to cell membranes that occurs during ferroptosis can be prevented by a protein called GPX4 acting together with a small

peptide called glutathione⁸, which confers antioxidant properties by combating lipid damage caused by peroxidation. As a consequence, approaches that inhibit GPX4 and that deplete glutathione from cells have recently been examined as potential anticancer strategies⁹. In 2016, researchers developed FIN56, a chemical that can induce ferroptosis, and this tool provided a clue that there might be another mechanism, besides GPX4- and glutathione-mediated repair, that protects against this process¹⁰. FIN56 was found to kill cells by perturbing a metabolic pathway called the mevalonate pathway, which acts upstream of ubiquinone synthesis; furthermore, supplementing cells with idebenone, a synthetic compound that is similar to ubiquinone, prevented FIN56 lethality, suggesting that FIN56 functions by depleting ubiquinone¹⁰. This previous work raised the question of whether ubiquinone normally functions to protect cells from ferroptosis. However, because FIN56 can deplete both ubiquinone and GPX4, it was difficult to draw any definitive conclusions.

Doll *et al.* and Bersuker *et al.* hypothesized that cells have a way of protecting themselves against ferroptosis even in the absence of GPX4. This was a heretical idea given that, during the short history of ferroptosis research, the dogma that GPX4 is essential to guard

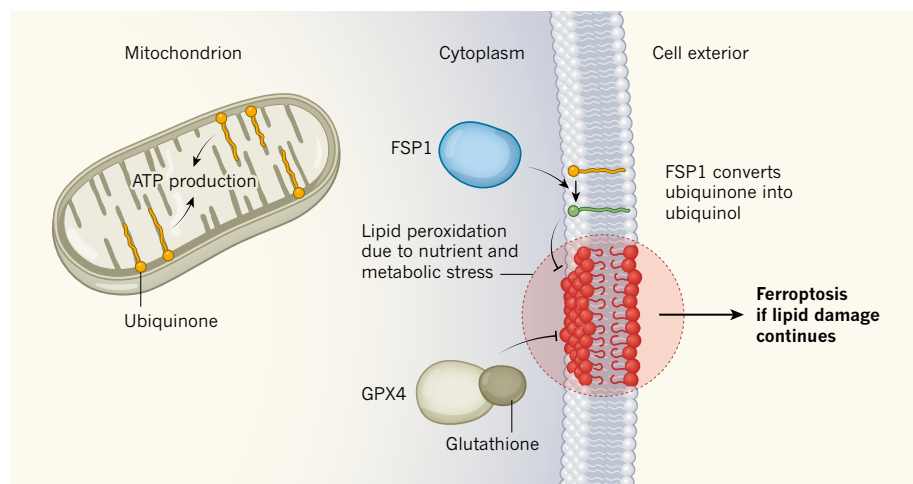


Figure 1 | A pathway that blocks cell death. Doll *et al.*² and Bersuker *et al.*³ report studies revealing that the FSP1 protein protects human cells from a type of cell death called ferroptosis. Some tumour cells are susceptible to ferroptosis, and this insight about FSP1 is of clinical interest. The finding has also uncovered a previously unknown role for a lipid called ubiquinone. Ubiquinone is found in lipid membranes, including those of an organelle called a mitochondrion, where it aids production of the molecule ATP, the cell's energy carrier¹. The authors report that FSP1 targets ubiquinone in the cell membrane to generate a reduced form of the molecule, called ubiquinol (green). Ferroptosis occurs if a form of lipid modification called peroxidation (red) damages the cell membrane; however, ubiquinol inhibits peroxidation and blocks ferroptosis. FSP1 acts independently of another pathway known to block lipid peroxidation and ferroptosis⁸, which requires the proteins GPX4 and glutathione.

against ferroptosis in all contexts had already been established. Nevertheless, these two teams searched for other such protective mechanisms. Both groups analysed human cells grown *in vitro* to test whether any components block ferroptosis when GPX4 is not present, and they independently identified a gene encoding a protein that they name ferroptosis suppressor protein 1 (FSP1), which was previously called AIFM2.

Excitingly, the authors discovered that FSP1 replenishes a reduced form of ubiquinone, called ubiquinol, that acts protectively by combating the lipid peroxidation that drives ferroptosis. Further experiments revealed that this FSP1-dependent modification of ubiquinone, in locations other than mitochondria, acts to protect against ferroptosis. The comic-book superhero Green Lantern has a power

ring that needs to be recharged once its protective energy becomes depleted, and, by analogy, FSP1's role in generating protective ubiquinol could be a similarly crucial recharging process.

The identification of this FSP1-mediated process suggests that drugs that inhibit FSP1 might be developed as anticancer treatments. Doll *et al.* and Bersuker *et al.* found that the level of resistance to ferroptosis across many human cancer cell lines grown *in vitro* correlates with the amount of FSP1 present in the cells, suggesting that modulating FSP1 might have clinical relevance. It would also be worth investigating whether treatments that boost FSP1 activity are useful as therapies for degenerative diseases driven by ferroptosis. These two latest studies clearly suggest that the mysteries of ferroptosis continue to yield important biological and therapeutic insights.

Brent R. Stockwell is in the Departments of Biological Sciences and of Chemistry, and at the Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York 10027, USA.

e-mail: bstockwell@columbia.edu

1. Crane, F. L. *Mitochondrion* **7** (Suppl.), S2–S7 (2007).
2. Doll, S. *et al.* *Nature* <https://doi.org/10.1038/s41586-019-1707-0> (2019).
3. Bersuker, K. *et al.* *Nature* <https://doi.org/10.1038/s41586-019-1705-2> (2019).
4. Dixon, S. J. *et al.* *Cell* **149**, 1060–1072 (2012).
5. Stockwell, B. R. & Jiang, X. *Cell Metab.* **30**, 14–15 (2019).
6. Hirschhorn, T. & Stockwell, B. R. *Free Radic. Biol. Med.* **133**, 130–143 (2019).
7. Stockwell, B. R. *et al.* *Cell* **171**, 273–285 (2017).
8. Yang, W. S. *et al.* *Cell* **156**, 317–331 (2014).
9. Liu, H., Schreiber, S. L. & Stockwell, B. R. *Biochemistry* **57**, 2059–2060 (2018).
10. Shimada, K. *et al.* *Nature Chem. Biol.* **12**, 497–503 (2016).

The author declares competing financial interests: see [go.nature.com/2bdv9eq](https://www.nature.com/2bdv9eq) for details.

two teams searched for other such protective mechanisms. Both groups analysed human cells grown *in vitro* to test whether any components block ferroptosis when GPX4 is not present, and they independently identified a gene encoding a protein that they name ferroptosis suppressor protein 1 (FSP1), which was previously called AIFM2.

Excitingly, the authors discovered that FSP1 replenishes a reduced form of ubiquinone, called ubiquinol, that acts protectively by combating the lipid peroxidation that drives ferroptosis. Further experiments revealed that this FSP1-dependent modification of ubiquinone, in locations other than mitochondria, acts to protect against ferroptosis. The comic-book superhero Green Lantern has a power ring that needs to be recharged once its protective energy becomes depleted, and, by analogy, FSP1's role in generating protective ubiquinol could be a similarly crucial recharging process.

The identification of this FSP1-mediated process suggests that drugs that inhibit FSP1 might be developed as anticancer treatments. Doll *et al.* and Bersuker *et al.* found that the level of resistance to ferroptosis across many human cancer cell lines grown *in vitro* correlates with the amount of FSP1 present in the cells, suggesting that modulating FSP1 might have clinical relevance. It would also be worth investigating whether treatments that boost FSP1 activity are useful as therapies for degenerative diseases driven by ferroptosis. These two latest studies clearly suggest that the mysteries of ferroptosis continue to yield important biological and therapeutic insights.

Brent R. Stockwell is in the Departments of Biological Sciences and of Chemistry, and at the Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York 10027, USA.
e-mail: bstockwell@columbia.edu

- Crane, F. L. *Mitochondrion* **7** (Suppl.), S2–S7 (2007).
- Doll, S. *et al.* *Nature* **575**, 693–698 (2019).
- Bersuker, K. *et al.* *Nature* **575**, 688–692 (2019).
- Dixon, S. J. *et al.* *Cell* **149**, 1060–1072 (2012).
- Stockwell, B. R. & Jiang, X. *Cell Metab.* **30**, 14–15 (2019).
- Hirschhorn, T. & Stockwell, B. R. *Free Radic. Biol. Med.* **133**, 130–143 (2019).
- Stockwell, B. R. *et al.* *Cell* **171**, 273–285 (2017).
- Yang, W. S. *et al.* *Cell* **156**, 317–331 (2014).
- Liu, H., Schreiber, S. L. & Stockwell, B. R. *Biochemistry* **57**, 2059–2060 (2018).
- Shimada, K. *et al.* *Nature Chem. Biol.* **12**, 497–503 (2016).

The author declares competing financial interests: see go.nature.com/2bdv9eq for details.

This article was published online on 21 October 2019.

Electrochemistry

Carbon dioxide efficiently converted to methanol

Xin-Ming Hu & Kim Daasbjerg

A molecular catalyst dispersed on carbon nanotubes has been found to catalyse the electrochemical conversion of carbon dioxide to methanol – a liquid fuel and industrially useful bulk chemical. **See p.639**

Molecular catalysts that mediate reactions with carbon dioxide often promote chemical reductions that form either carbon monoxide or formic acid (HCO₂H), but lack the activity and selectivity to reduce these compounds further to make other useful products, such as methanol, ethanol or methane. On page 639, Wu *et al.*¹ report that a molecular catalyst immobilized on carbon nanotubes can promote the electrochemical conversion of CO₂ to methanol in water. The result holds promise for advancing the search for catalysts that make highly reduced products from CO₂. Such products can then be used as fuels and as feedstock chemicals for industrial processes.

The heavy use of fossil fuels has led to excessive emissions of CO₂ into the atmosphere, and poses imminent threats to our climate system. Renewable energy sources, such as solar and wind power, are green and sustainable alternatives to fossil fuels for powering our society. Unfortunately, their intermittent nature limits their widespread use. Methods for storing the energy from these sources are therefore needed, to even out the supply.

Electrically powered methods for transforming CO₂ into fuels and other CO₂-derived chemicals are a promising strategy for tackling some of these energy issues, with the added bonus that they might help to mitigate atmospheric CO₂ levels². But the development of such methods is by no means easy, because CO₂ is a stable and relatively unreactive molecule. Catalysts are therefore essential to activate CO₂ and drive its conversion into desired products.

Among the various classes of catalytic material, molecular catalysts that consist of ligand molecules bound to metal ions have certain advantages: they follow well-characterized reaction pathways, and have chemically modifiable structures that allow their activity to be tuned quite precisely. Molecular catalysts that promote the electrochemical conversion of CO₂ to carbon monoxide or, in fewer cases, to formic acid (or its formate salt) have been known for decades³. Attempts to reduce these compounds further to make methanol, ethanol, methane or ethylene (CH₂=CH₂) have been unsuccessful – or, at best, have provided

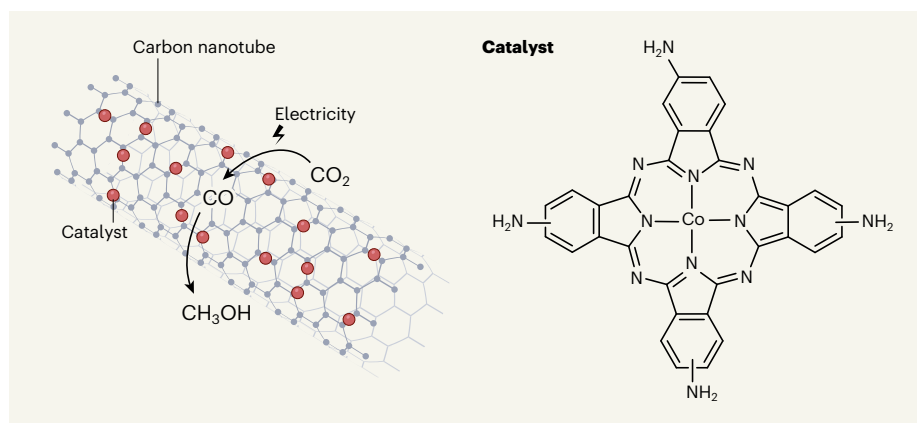


Figure 1 | Electrochemical production of methanol from carbon dioxide. Wu *et al.*¹ report that a cobalt phthalocyanine catalyst immobilized on carbon nanotubes can electrochemically reduce carbon dioxide in water. The reaction first produces carbon monoxide, which is reduced further to methanol (CH₃OH), an important liquid fuel and bulk chemical. The conversion of CO₂ to methanol using molecular catalysts has previously been ineffective. The point of attachment of three of the amino (NH₂) groups to the benzene rings in the catalyst is not known. Co, cobalt.

these products in small quantities with low selectivity^{4,5} (that is, as a small component of a mixture with other products). Copper-based materials have previously been the most successful catalysts for such reactions⁶.

Wu *et al.* now reveal that, when a complex called cobalt phthalocyanine is dispersed on carbon nanotubes, it has appreciable catalytic activity and selectivity for the electrochemical reduction of CO₂ to methanol (Fig. 1). More specifically, the cobalt phthalocyanine complex must be physically adsorbed to the surface of the carbon nanotubes as individual molecules. The key finding is that this mixed catalyst system not only activates CO₂ to produce carbon monoxide, but also, surprisingly, promotes further reduction to methanol when high voltages are applied in the electrochemical cell.

The researchers found that optimization of the catalytic system was difficult, because many extrinsic factors affected the activity of the molecular catalyst. These included the method used to immobilize the catalyst on the support; the specific carbon support chosen; the ratio of the concentration of the catalyst to that of the support; and the voltage used for the electrochemical reduction. The product selectivity of CO₂ reductions catalysed by cobalt phthalocyanine can be strongly affected by even a subtle variation in any of these factors⁷. However, the optimized catalyst system has significantly improved activity and selectivity compared with previous molecular-catalyst systems. Still, it is not as good as the state-of-the-art, solid-state metallic catalysts that have been reported for methanol production^{8,9}.

A long-standing issue associated with molecular catalysts in general is their long-term stability. Wu *et al.* found that their cobalt phthalocyanine system lost its catalytic activity over the course of five hours, and they identified the deactivation process as degradation of the phthalocyanine ligand. When they modified the ligand by appending amino (NH₂) substituents to it, they found that their system's stability was enhanced – it lasted for more than 12 hours, with only a slight loss of overall activity and selectivity. The reason for the stabilizing effect is not known. Note, however, that the catalyst would need to last for thousands of hours if the reduction process were to be implemented in an industrial setting.

The findings reveal that molecular catalysts have great prospects for use in CO₂ transformations. Future research could focus on further improving the activity, selectivity and stability of the molecular catalyst–carbon nanotube hybrid system through judicious chemical manipulations of the catalyst and the support, and of the interactions between them. Detailed mechanistic insight into the catalytic conversion of CO₂ to carbon monoxide, and

further to methanol, might be gained using computational modelling and ‘operando’ characterization techniques, which monitor the consumption of reactants and the build-up of products during catalysis. Such efforts would lay the foundations not only for improving the performance of existing systems, but also for discovering new catalysts involving metal complexes, or structurally similar catalysts consisting of single metal atoms dispersed in carbon materials¹⁰.

Concerns have been raised that the generally moderate activity, selectivity and stability of molecular catalysts for CO₂ reactions will prevent them from being used on an industrial scale. Moreover, the transport of CO₂ in electrochemical cells that have been used in proof-of-concept experiments is limited by the low solubility of this gas in water¹¹. However, the adoption of flow technology in which a large quantity of gaseous CO₂ is fed directly to catalysts can greatly improve the outcome of CO₂ transformations¹². With continued efforts to improve catalyst performance and the design of electrochemical cells, the

industrial production of methanol from CO₂ could well be within reach.

Xin-Ming Hu and **Kim Daasbjerg** are in the Carbon Dioxide Activation Center, Interdisciplinary Nanoscience Center and Department of Chemistry, Aarhus University, 8000 Aarhus, Denmark.
e-mail: kdaa@chem.au.dk

1. Wu, Y., Jiang, Z., Lu, X., Liang, Y. & Wang, H. *Nature* **575**, 639–642 (2019).
2. Whipple, D. T. & Kenis, P. J. A. *J. Phys. Chem. Lett.* **1**, 3451–3458 (2010).
3. Elouarzaki, K., Kannan, V., Jose, V., Sabharwal, H. S. & Lee, J.-M. *Adv. Energy Mater.* **9**, 1900090 (2019).
4. Kapusta, S. & Hackerman, N. *J. Electrochem. Soc.* **131**, 1511–1514 (1984).
5. Shen, J. *et al. Nature Commun.* **6**, 8177 (2015).
6. Nitopi, S. *et al. Chem. Rev.* **119**, 7610–7672 (2019).
7. Boutin, E. *et al. Angew. Chem. Int. Edn* **58**, 16172–16176 (2019).
8. Lu, L. *et al. Angew. Chem. Int. Edn* **57**, 14149–14153 (2018).
9. Zhang, W. *et al. Angew. Chem. Int. Edn* **57**, 9475–9479 (2018).
10. Peng, Y., Lu, B. & Chen, S. *Adv. Mater.* **30**, 1801995 (2018).
11. Weekes, D. M., Salvatore, D. A., Reyes, A., Huang, A. & Berlinguette, C. P. *Acc. Chem. Res.* **51**, 910–918 (2018).
12. Burdyny, T. & Smith, W. A. *Energy Environ. Sci.* **12**, 1442–1453 (2019).

Microbiology

Fresh ammunition in bacterial warfare

Brent W. Anderson & Jue D. Wang

A previously unknown bacterial toxin has now been characterized. The protein is secreted into neighbouring cells, depleting them of essential energy-carrying molecules and so leading to the cells' demise. **See p.674**

To survive, bacteria must monopolize valuable resources. One way to do this is to attack and outcompete neighbouring cells – for example using the type VI secretion system, which injects neighbours with a toxin that can inhibit their growth or kill them¹. On page 674, Ahmad *et al.*² describe a previously unknown toxin, TasI, used in the type VI secretion system of the pathogen *Pseudomonas aeruginosa*. TasI launches a two-pronged attack on cells: not only does it rapidly deplete them of essential energy-carrying ATP molecules, but it also produces a signalling molecule that prevents the synthesis of more ATP.

Ahmad *et al.* made their discovery when studying a highly virulent strain of *P. aeruginosa*. The authors identified a region of the bacterium's genome that encodes a protein allowing *P. aeruginosa* to outcompete other bacteria. The amino-acid sequence of this toxin had no obvious similarity to any other proteins secreted by the type VI system.

The authors found that the toxin was structurally similar to a class of enzyme that synthesizes the ‘alarmone’ molecules guanosine tetraphosphate (ppGpp) and guanosine pentaphosphate (pppGpp), collectively referred to as (p)ppGpp. Alarmones are signalling molecules produced by bacteria and plants to help them to survive stressful conditions. Production of (p)ppGpp is a near-universal response to stresses such as nutrient starvation in bacteria. Its production causes a decrease in bacterial growth³, preventing excessive proliferation and so allowing bacteria to survive in low-nutrient conditions.

It seems logical for a bacterial toxin to produce (p)ppGpp as a way of slowing the growth of competitor cells, so Ahmad and colleagues tested the enzymatic capabilities of the purified *P. aeruginosa* toxin. Unexpectedly, the protein did not produce (p)ppGpp. Instead, it produced the related alarmone (p)ppApp, which comprises adenosine tetraphosphate

News & views

Microbiology

Fresh ammunition in bacterial warfare

Brent W. Anderson & Jue D. Wang

A previously unknown bacterial toxin has now been characterized. The protein is secreted into neighbouring cells, depleting them of essential energy-carrying molecules and so leading to the cells' demise.

To survive, bacteria must monopolize valuable resources. One way to do this is to attack and outcompete neighbouring cells – for example using the type VI secretion system, which injects neighbours with a toxin that can inhibit their growth or kill them¹. Writing in *Nature*, Ahmad *et al.*² describe a previously unknown toxin, Tas1, used in the type VI secretion system of the pathogen *Pseudomonas aeruginosa*. Tas1 launches a two-pronged attack on cells: not only does it rapidly deplete them of essential energy-carrying ATP molecules, but it also produces a signalling molecule that prevents the synthesis of more ATP.

Ahmad *et al.* made their discovery when studying a highly virulent strain of *P. aeruginosa*. The authors identified a region of the bacterium's genome that encodes a protein allowing *P. aeruginosa* to outcompete other bacteria. The amino-acid sequence of this toxin had no obvious similarity to any other proteins secreted by the type VI system.

The authors found that the toxin was structurally similar to a class of enzyme that synthesizes the 'alarmone' molecules guanosine tetraphosphate (ppGpp) and guanosine pentaphosphate (pppGpp), collectively referred to as (p)ppGpp. Alarmones are signalling molecules produced by bacteria and plants to help them to survive stressful conditions. Production of (p)ppGpp is a near-universal response to stresses such as nutrient starvation in bacteria. Its production causes a decrease in bacterial growth³, preventing excessive proliferation and so allowing bacteria to survive in low-nutrient conditions.

It seems logical for a bacterial toxin to produce (p)ppGpp as a way of slowing the growth of competitor cells, so Ahmad and colleagues tested the enzymatic capabilities of the purified *P. aeruginosa* toxin. Unexpectedly,

the protein did not produce (p)ppGpp. Instead, it produced the related alarmone (p)ppApp, which comprises adenosine tetraphosphate (ppApp) and adenosine pentaphosphate (pppApp) molecules. The authors therefore named the toxin type VI secretion effector (p)ppApp synthetase I, or Tas1 for short. This is the first example of an alarmone-producing enzyme being transported between bacteria – a remarkable fact, given that these enzymes are found in nearly all bacteria.

Type VI systems often secrete enzymes that destroy essential cellular structures, such as the cell wall, the cell membrane or the genome itself⁴. But Ahmad *et al.* found that the toxicity of Tas1 is linked to its synthesis of (p)ppApp from ATP (Fig. 1). ATP is crucial for almost every cellular process, from DNA replication to the production of proteins and maintenance of the cell's structural integrity. Tas1 synthesizes (p)ppApp from ATP strikingly quickly – one molecule of toxin produces 180,000 (p)ppApp

molecules per minute. At such a rate, the toxin depletes the target cell of ATP within minutes, simultaneously disrupting several essential metabolic pathways. Of note, *P. aeruginosa* secretes other toxins alongside Tas1, some of which attack cellular structures that require ATP for their synthesis, including the cell wall and membrane. Tas1 activity might therefore compound the effects of these other toxins.

Ahmad and colleagues went on to highlight the toxic role of (p)ppApp in influencing bacterial physiology. Little has been reported about how (p)ppApp is produced and what it does in bacteria⁵. The authors found that (p)ppApp blocks ATP synthesis in the target cell by binding and inhibiting PurF, a key enzyme in the synthesizing process. Thus, (p)ppApp probably prevents the cell from regenerating ATP and so escaping the death spiral induced by the alarmone's own production. More work is needed to delineate how much this role for (p)ppApp contributes to the overall toxicity of Tas1 in target cells.

Alarmone production is highly regulated to ensure that the molecules are synthesized only when needed, and degraded when stress has passed. Tas1 production of (p)ppApp overrides these rules – (p)ppApp is synthesized with abandon and, as the authors show, there are unlikely to be any enzymes that can degrade (p)ppApp quickly enough to avoid cell death. Nonetheless, this newfound understanding of (p)ppApp can augment our knowledge of other alarmones. (p)ppGpp, which is structurally similar to (p)ppApp, controls cell growth in part by inhibiting proteins involved in the synthesis of energy-carrying molecules such as ATP and GTP^{3,6–8}, including PurF. The fact that both (p)ppApp and (p)ppGpp inhibit this protein, along with the structural similarity

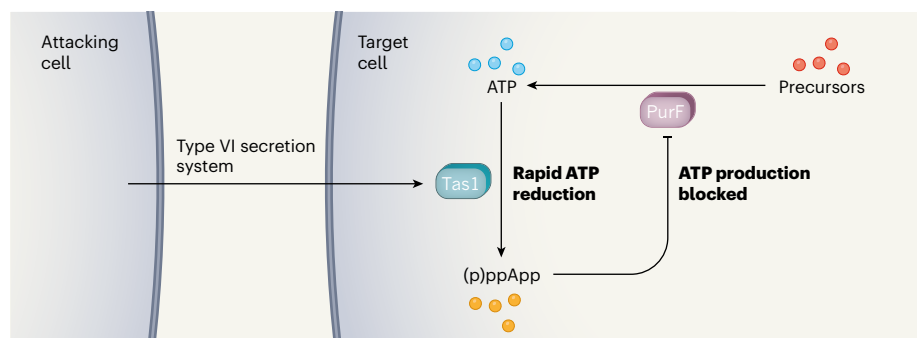


Figure 1 | A two-pronged attack system. Bacteria can attack target cells using cellular machinery called the type VI secretion system. Ahmad *et al.*² find that the type VI system of one bacterium, *Pseudomonas aeruginosa*, secretes a previously unknown toxin, which the authors name Tas1. Tas1 uses energy-carrying ATP molecules to produce the signalling molecule (p)ppApp, rapidly reducing ATP levels. In turn, (p)ppApp blocks production of ATP by inhibiting the first enzyme in the ATP-synthesis pathway, PurF. This two-pronged attack depletes target cells of essential ATP within minutes, causing death.

between the two alarmones, led Ahmad *et al.* to hypothesize that the molecules could have many overlapping targets.

TasI is the only dedicated (p)ppApp-synthesizing enzyme found so far. However, (p)ppApp has been detected in some bacteria, in which its physiological role has yet to be determined⁸. Clearly, it is unlikely to act as a toxin in these cells. Ahmad and colleagues' discovery that (p)ppApp inhibits PurF is the first step towards mapping the network of targets regulated by this alarmone in healthy cells. Doing so should help us to gain a broader understanding of how alarmone regulatory pathways rewire bacterial physiology.

Type VI secretion systems provide bacteria

with weapons against competitors, increasing their ability to thrive in a range of environments – from plants to the human intestinal tract to hospitals^{9,10}. The discovery of a toxin that so irreversibly suppresses competitor metabolism opens a new chapter in our understanding of the ammunition used in interbacterial warfare. It will be exciting to see whether other examples of this toxin are found across the bacterial domain, or perhaps even in bacterium–host interactions.

Brent W. Anderson and Jue D. Wang

are in the Department of Bacteriology, University of Wisconsin–Madison, Wisconsin 53706, USA.

e-mails: bwanderson3@wisc.edu; wang@bact.wisc.edu

1. Coulthurst, S. *Microbiology* **165**, 503–515 (2019).
2. Ahmad, S. *et al.* *Nature* <https://doi.org/10.1038/s41586-019-1735-9> (2019).
3. Gourse, R. L. *et al.* *Annu. Rev. Microbiol.* **72**, 163–184 (2018).
4. Durand, E., Cambillau, C., Cascales, E. & Journet, L. *Trends Microbiol.* **22**, 498–507 (2014).
5. Sobala, M., Bruhn-Olszewska, B., Cashel, M. & Potrykus, K. *Front. Microbiol.* **10**, 859 (2019).
6. Liu, K. *et al.* *Mol. Cell* **57**, 735–749 (2015).
7. Wang, B. *et al.* *Nature Chem. Biol.* **15**, 141–150 (2019).
8. Zhang, Y., Zborníková, E., Rejman, D. & Gerdes, K. *mBio* **9**, e02188-17 (2018).
9. Coyne, M. J. & Comstock, L. E. *Microbiol. Spectrum* **7**, <https://doi.org/10.1128/microbiolspec.PSIB-0009-2018> (2019).
10. Bernal, P., Llamas, M. A. & Filloux, A. *Environ. Microbiol.* **20**, 1–15 (2018).

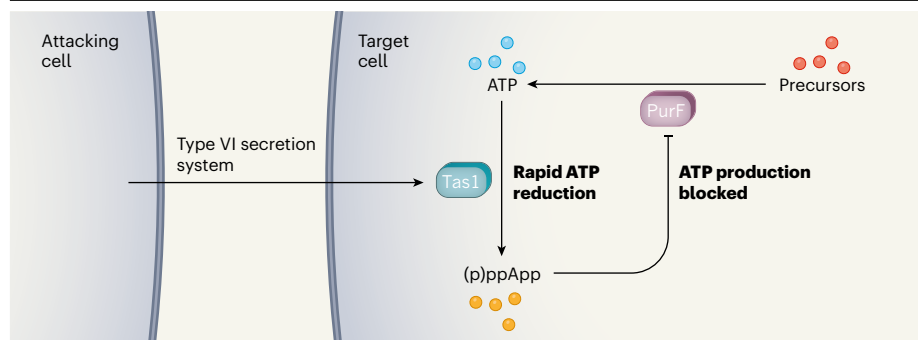


Figure 1 | A two-pronged attack system. Bacteria can attack target cells using cellular machinery called the type VI secretion system. Ahmad *et al.*² find that the type VI system of one bacterium, *Pseudomonas aeruginosa*, secretes a previously unknown toxin, which the authors name Tas1. Tas1 uses energy-carrying ATP molecules to produce the signalling molecule (p)ppApp, rapidly reducing ATP levels. In turn, (p)ppApp blocks production of ATP by inhibiting the first enzyme in the ATP-synthesis pathway, PurF. This two-pronged attack depletes target cells of essential ATP within minutes, causing death.

(ppApp) and adenosine pentaphosphate (pppApp) molecules. The authors therefore named the toxin type VI secretion effector (p)ppApp synthetase 1, or Tas1 for short. This is the first example of an alarmone-producing enzyme being transported between bacteria – a remarkable fact, given that these enzymes are found in nearly all bacteria.

Type VI systems often secrete enzymes that destroy essential cellular structures, such as the cell wall, the cell membrane or the genome itself⁴. But Ahmad *et al.* found that the toxicity of Tas1 is linked to its synthesis of (p)ppApp from ATP (Fig. 1). ATP is crucial for almost every cellular process, from DNA replication to the production of proteins and maintenance of the cell's structural integrity. Tas1 synthesizes (p)ppApp from ATP strikingly quickly – one molecule of toxin produces 180,000 (p)ppApp molecules per minute. At such a rate, the toxin depletes the target cell of ATP within minutes, simultaneously disrupting several essential metabolic pathways. Of note, *P. aeruginosa* secretes other toxins alongside Tas1, some of which attack cellular structures that require ATP for their synthesis, including the cell wall and membrane. Tas1 activity might therefore compound the effects of these other toxins.

Ahmad and colleagues went on to highlight the toxic role of (p)ppApp in influencing bacterial physiology. Little has been reported about how (p)ppApp is produced and what it does in bacteria⁵. The authors found that (p)ppApp blocks ATP synthesis in the target cell by binding and inhibiting PurF, a key enzyme in the synthesizing process. Thus, (p)ppApp probably prevents the cell from regenerating ATP and so escaping the death spiral induced by the alarmone's own production. More work is needed to delineate how much this role for (p)ppApp contributes to the overall toxicity of Tas1 in target cells.

Alarmone production is highly regulated to ensure that the molecules are synthesized only when needed, and degraded when stress has passed. Tas1 production of (p)ppApp overrides

these rules – (p)ppApp is synthesized with abandon and, as the authors show, there are unlikely to be any enzymes that can degrade (p)ppApp quickly enough to avoid cell death. Nonetheless, this newfound understanding of (p)ppApp can augment our knowledge of other alarmones. (p)ppGpp, which is structurally similar to (p)ppApp, controls cell growth in part by inhibiting proteins involved in the synthesis of energy-carrying molecules such as ATP and GTP^{3,6–8}, including PurF. The fact that both (p)ppApp and (p)ppGpp inhibit this protein, along with the structural similarity between the two alarmones, led Ahmad *et al.* to hypothesize that the molecules could have many overlapping targets.

Tas1 is the only dedicated (p)ppApp-synthesizing enzyme found so far. However, (p)ppApp has been detected in some bacteria, in which its physiological role has yet to be determined⁸. Clearly, it is unlikely to act as

a toxin in these cells. Ahmad and colleagues' discovery that (p)ppApp inhibits PurF is the first step towards mapping the network of targets regulated by this alarmone in healthy cells. Doing so should help us to gain a broader understanding of how alarmone regulatory pathways rewire bacterial physiology.

Type VI secretion systems provide bacteria with weapons against competitors, increasing their ability to thrive in a range of environments – from plants to the human intestinal tract to hospitals^{9,10}. The discovery of a toxin that so irreversibly suppresses competitor metabolism opens a new chapter in our understanding of the ammunition used in interbacterial warfare. It will be exciting to see whether other examples of this toxin are found across the bacterial domain, or perhaps even in bacterium–host interactions.

Brent W. Anderson and Jue D. Wang

are in the Department of Bacteriology, University of Wisconsin–Madison, Wisconsin 53706, USA.

e-mails: bwanderson3@wisc.edu; wang@bact.wisc.edu

1. Coulthurst, S. *Microbiology* **165**, 503–515 (2019).
2. Ahmad, S. *et al.* *Nature* **575**, 674–678 (2019).
3. Gourse, R. L. *et al.* *Annu. Rev. Microbiol.* **72**, 163–184 (2018).
4. Durand, E., Cambillau, C., Cascales, E. & Journet, L. *Trends Microbiol.* **22**, 498–507 (2014).
5. Sobala, M., Bruhn-Olszewska, B., Cashel, M. & Potrykus, K. *Front. Microbiol.* **10**, 859 (2019).
6. Liu, K. *et al.* *Mol. Cell* **57**, 735–749 (2015).
7. Wang, B. *et al.* *Nature Chem. Biol.* **15**, 141–150 (2019).
8. Zhang, Y., Zborniková, E., Rejman, D. & Gerdes, K. *mBio* **9**, e02188–17 (2018).
9. Coyne, M. J. & Comstock, L. E. *Microbiol. Spectrum* **7**, <https://doi.org/10.1128/microbiolspec.PSIB-0009-2018> (2019).
10. Bernal, P., Llamas, M. A. & Filloux, A. *Environ. Microbiol.* **20**, 1–15 (2018).

This article was published online on 6 November 2019.

Solid-state physics

Surface polarization feels the heat

Gustau Catalan & Beatriz Noheda

A crystal's surface has been found to behave as a distinct material that has temperature-dependent electrical polarization – despite the rest of the crystal being non-polar.

When crystals of certain materials are squeezed, the compression causes a separation of internal charge – a polarization – that generates a voltage. This phenomenon is known as piezoelectricity. Some piezoelectric materials also exhibit spontaneous polarization that changes in magnitude with

increasing temperature. These materials are said to be pyroelectric, and are useful in heat sensors and for solid-state cooling (because pyroelectrics change temperature in an applied electric field)¹. Pyroelectrics have thus been intensively investigated, with research naturally focusing on electrically

polar materials. Writing in *Advanced Materials*, however, Meirzadeh *et al.*² report that the non-polar material strontium titanate (SrTiO_3) is also pyroelectric, suggesting that the net needs to be cast more widely in the search for pyroelectrics.

Conventional piezo- and pyroelectricity ultimately arise from the fact that the repeating unit (the unit cell) of the crystal lattice is asymmetrical. A perfect, infinite crystal of strontium titanate is symmetrical and therefore should not be pyroelectric. But perfection, alas, does not exist. Many crystals contain defects whose concentration varies across the crystal; the resulting concentration gradient breaks the macroscopic symmetry of the crystal, causing residual piezoelectricity and pyroelectricity³.

Moreover, even the most perfect crystals are finite, which means that they inevitably have one kind of ‘defect’: surfaces. And surfaces break symmetry, because what is above the surface is different from what is below. Hence, irrespective of the intrinsic symmetry of the bulk, surfaces can, in theory, be polar and even pyroelectric. This seems to be the case for strontium titanate, a cubic crystal commonly used as a substrate for growing films of other oxides.

Determining whether pyroelectricity comes from the surface, rather than from inside a crystal, is not trivial. Meirzadeh and co-workers did so by heating the surface of strontium titanate with fast laser pulses, and measuring how the resulting pyroelectric current evolves with time (Fig. 1). The rate at which the current decays is related to the rate at which the surface reaches thermal

equilibrium, a process called thermalization: fast decay of the current implies quick thermalization and therefore suggests that the depth of the pyroelectric region is shallow.

From the time-dependence of the signal, the authors estimate that the depth of the polarized layer in strontium titanate is about 1.2 nanometres, equivalent to 3 unit cells. This coincides with an intrinsic region of polar distortion that has been predicted by first-principles calculations to form at the surface of strontium titanate as a result of surface tension^{2,4}. Therefore, the pyroelectricity seems to arise from an inherent surface distortion.

The authors took precautions to discard alternative explanations: they checked that the direction of the heat-induced current does not depend on the orientation of the crystal, ruling out a bulk effect; and that the local heating produced by the laser is very small (the temperature increases are at the sub-kelvin scale), which means that the strain gradients induced by thermal expansion are insignificant. Other experiments and data analysis were carried out to exclude the possibility that the induced current is due to molecules (typically water) adsorbed to the surface, charges trapped by lattice defects, excitation of free electrons induced by light, or the thermoelectric Seebeck effect (which generates currents in semiconductors that contain temperature gradients). Importantly, the pyroelectricity disappeared when Meirzadeh *et al.* deposited an atomically thin layer of amorphous silica (SiO_2) on top of the strontium titanate, consistent with the idea that the phenomenon originates at the surface.

Moreover, the temperature dependence of the surface polarization suggests that a phase transition occurs that is not observed in the bulk. This is interesting, because it implies that the pyroelectricity does not simply arise from thermal expansion of the piezoelectric surface⁵, but from a true phase transition confined to the surface.

Surface layers of crystals known as skin layers, which have different properties from those of the bulk, are found in various materials^{6,7}, including strontium titanate⁸. However, such skin layers tend to be much thicker than the atomically thin one described by Meirzadeh and colleagues, and are probably induced by defects introduced during polishing, rather than being intrinsic. Rearrangements of surface atoms in strontium titanate have also previously been reported⁹, but it has not been established whether the resulting surfaces are pyroelectric. Meirzadeh and colleagues’ findings are therefore new.

This discovery matters for many reasons. One is pointed out by the authors: multilayered thin-film devices could be designed to take advantage of the surface polarization at the interface between each layer^{10,11}. There are also consequences for bulk crystals. When a crystal of any symmetry is bent, it can become electrically polarized as a result of strain being produced non-uniformly in the material – a phenomenon called flexoelectricity. If the surfaces are already polar, then the surface polarization will also contribute to the total flexoelectricity^{12,13}. In fact, the surface termination of a strontium titanate crystal (that is, whether the last atomic layer is TiO_2 or SrO) can theoretically change the sign of the flexoelectric voltage – even for macroscopic crystals¹⁴.

Surfaces are also interesting in themselves, being 2D entities in a 3D world. If a pyroelectric phase transition does occur in the surface of strontium titanate, it would offer an excellent playground for testing models of the effects of dimensionality on phase transitions in general, because of the universality of the laws that underpin such transitions¹⁵. It will also be interesting to study the nature of the dipoles that form at surfaces and, specifically, whether their orientation can be switched by an applied voltage – in other words, whether the surface of strontium titanate is not just a 2D pyroelectric but also a 2D ferroelectric.

Electrical polarity might not be the only surprising thing about the surface of strontium titanate. Although this material is an insulator, its surface conducts electricity¹⁶. The surface might therefore be a polar metal: an exotic type of metal that contains electric dipoles^{17,18}. Polar metals have been much sought after, partly out of fundamental curiosity (polar materials are normally insulators

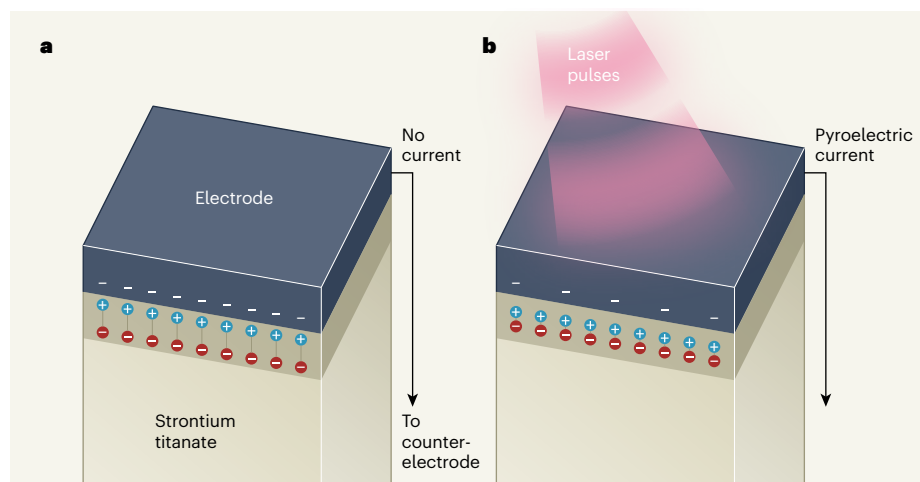


Figure 1 | Pyroelectricity at the surface of strontium titanate. Meirzadeh *et al.*² report that the surface of crystals of strontium titanate undergoes temperature-dependent changes of electrical polarization – a phenomenon known as pyroelectricity. **a**, In the authors’ experiments, the surface layers (darker tint) have an initial amount of intrinsic polarization, which is balanced (screened) by free charges in an overlying electrode. **b**, Laser light heats up the surface and lowers the polarization (reduces the distance between charges in the dipoles). This causes a current to flow, to balance out the modified polarization. The current disappears once the temperature stabilizes; the time taken for this to happen carries information about the time taken to reach thermal equilibrium, which is proportional to the volume of pyroelectric material. The authors thus find that the volume is very small, consistent with a thin surface layer.

or, at most, semiconductors), but also because they are expected to have unique electronic properties¹⁹. Meirzadeh and colleagues' findings hint that polar metals might have been under our noses all along, paradoxically on the surfaces of non-polar insulators.

Gustau Catalan is at Institutio Catalana de Recerca i Estudis Avançats, Barcelona 08010, Spain, and at Institut Català de Nanociència i Nanotecnologia, CSIC-BIST, Barcelona.

Beatriz Noheda is at the Zernike Institute for Advanced Materials, University of Groningen, Groningen 9747AG, the Netherlands.
e-mails: gustau.catalan@icn2.cat;
b.noheda@rug.nl

1. Whatmore, R. W. *Rep. Prog. Phys.* **49**, 1335–1386 (1986).
2. Meirzadeh, E. *et al. Adv. Mater.* **31**, 1904733 (2019).
3. Biancoli, A., Fancher, C. M., Jones, J. L. & Damjanovic, D. *Nature Mater.* **14**, 224–229 (2015).
4. Padilla, J. & Vanderbilt, D. *Surf. Sci.* **418**, 64–70 (1998).

5. Bhalla, A. S. & Newnham, R. E. *Phys. Status Solidi A* **58**, K19–K24 (1980).
6. Gehring, P. M., Hirota, K., Majkrzak, C. F. & Shirane, G. *Phys. Rev. Lett.* **71**, 1087 (1993).
7. Domingo, N., Bagués, N., Santiso, J. & Catalan, G. *Phys. Rev. B* **91**, 094111 (2015).
8. Hirota, K., Hill, J. P., Shapiro, S. M., Shirane, G. & Fujii, Y. *Phys. Rev. B* **52**, 13195–13205 (1995).
9. Jiang, Q. & Zegenhagen, J. *Surf. Sci.* **367**, L42–L46 (1996).
10. Sai, N., Meyer, B. & Vanderbilt, D. *Phys. Rev. Lett.* **84**, 5636 (2000).
11. Yamada, H., Kawasaki, M., Ogawa, Y. & Tokura, Y. *Appl. Phys. Lett.* **81**, 4793 (2002).
12. Tagantsev, K. *Phys. Rev. B* **34**, 5883–5889 (1986).
13. Narvaez, J., Vasquez-Sancho, F. & Catalan, G. *Nature* **538**, 219–221 (2016).
14. Stengel, M. *Phys. Rev. B* **90**, 201112(R) (2014).
15. Stanley, H. E. *Rev. Mod. Phys.* **71**, S358 (1999).
16. Santander-Syro, A. F. *et al. Nature* **469**, 189–193 (2011).
17. Anderson, P. W. & Blount, E. I. *Phys. Rev. Lett.* **14**, 217–219 (1965).
18. Shi, Y. *et al. Nature Mater.* **12**, 1024–1027 (2013).
19. Benedek, N. A. & Biro, T. *J. Mater. Chem. C* **4**, 4000–4015 (2016).

This article was published online on 18 November 2019.

Evolution

The balancing act of growth and expansion

Henry Mattingly & Thierry Emonet

Bacteria move along gradients of chemical attractants. Two studies find that, in nutrient-rich environments, bacteria can grow rapidly by following a non-nutritious attractant – but expanding too fast leaves them vulnerable. See p.658 & p.664

Bacteria can sense chemical attractants and use that information to navigate towards resources or away from harm – a process called chemotaxis. But why bacteria chase signals that often do not have much nutritional value has been a long-standing puzzle. Cremer *et al.*¹ show on page 658 that bacterial populations can use non-nutritious attractants as cues for rapidly expanding through nutrient-rich areas, ensuring that plentiful nutrients are available for their future growth. And on page 664, Liu *et al.*² build on this work to reveal an unanticipated rule of bacterial evolution: the safest way for a bacterial population to colonize a habitat is not necessarily to expand as fast as possible, because rapid expansion can leave the population vulnerable to invasion by competitors.

In the 1960s, the biochemist Julius Adler demonstrated that a group of cells consuming a chemical attractant can form a rapidly expanding wave that follows a moving concentration gradient that the cells create on their own³. That is, by consuming the attractant in their immediate vicinity, the cells create a gradient between their current location and the

surrounding regions in which the chemical has not yet been consumed. The cells then chase the higher concentration – rather like a horse chasing a carrot on a stick. The wave's

expansion speed is determined by how fast the travelling cells deplete the local attractant⁴.

Cremer *et al.* examined how a cell population's use of chemotaxis to expand (defined as the occupation of more space), as in Adler's experiments, affects its growth (the increase in cell numbers). The authors seeded small colonies of bacteria in a Petri dish, and measured population size over time as the cells grew and filled the available space. As Adler had observed, the colonies formed expanding waves, and some cells fell behind the wavefront, seeding the newly covered ground.

Importantly, when Cremer and colleagues added small amounts of a non-nutritious chemical attractant that was different from the nutrient on which the cells were growing, the population capitalized on chemotaxis to expand before the local nutrient had become depleted. This increased the number of cells that had access to nutrients at a given time and allowed the population to grow much faster than it did without the directional cue of the attractant (Fig. 1). This gain relied on a separation between chemotaxis and growth: the attractant served as a cue, rather than as a nutrient source⁵, to direct the cells towards unoccupied territory. When the attractant is the only nutrient, the population does not grow as fast; either the attractant is abundant, and the cells can't consume it fast enough for rapid expansion, or the attractant is limited, and expansion can be fast but the settlers behind the wavefront are starved and don't grow.

This work demonstrates that – in a nutrient-rich environment – the faster a single population expands, the faster it grows. But what happens when competitors (including spontaneously generated mutants in the population) expand into the same territory? Last year, we and our colleagues⁶ showed that bacteria with different chemotactic abilities

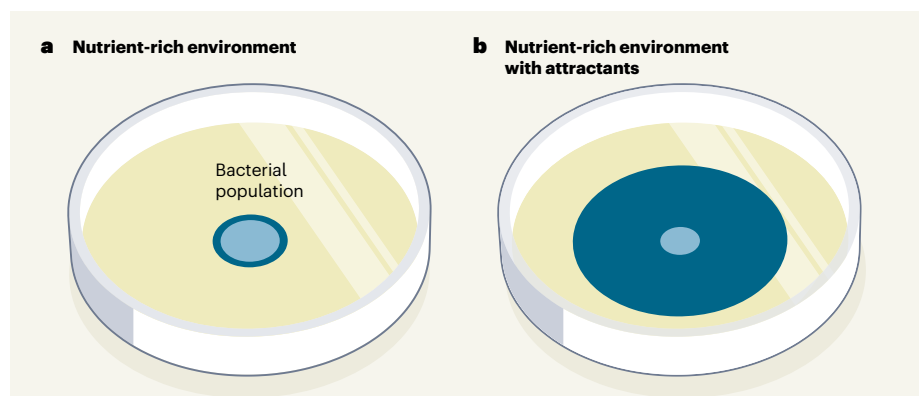


Figure 1 | How bacteria maximize growth in nutrient-rich environments. **a**, Populations of bacteria can spread out within a nutrient-rich habitat through cell division and random motion. But this approach causes most of the population to stay in a small location and deplete local nutrients – so, many cells starve (light blue) and only the outer edge grows (dark blue). **b**, Cremer *et al.*¹ grew bacteria in the same environment but included low levels of a non-nutritious attractant chemical (not shown). The cells chased the attractant through a process called chemotaxis, expanding rapidly across the dish before the local nutrients were depleted, so that most of the population had the nutrients needed to grow.

or, at most, semiconductors), but also because they are expected to have unique electronic properties¹⁹. Meirzadeh and colleagues' findings hint that polar metals might have been under our noses all along, paradoxically on the surfaces of non-polar insulators.

Gustau Catalan is at Institutio Catalana de Recerca i Estudis Avançats, Barcelona 08010, Spain, and at Institut Català de Nanociència i Nanotecnologia, CSIC-BIST, Barcelona.

Beatriz Noheda is at the Zernike Institute for Advanced Materials, University of Groningen, Groningen 9747AG, the Netherlands.
e-mails: gustau.catalan@icn2.cat;
b.noheda@rug.nl

1. Whatmore, R. W. *Rep. Prog. Phys.* **49**, 1335–1386 (1986).
2. Meirzadeh, E. *et al. Adv. Mater.* **31**, 1904733 (2019).
3. Biancoli, A., Fancher, C. M., Jones, J. L. & Damjanovic, D. *Nature Mater.* **14**, 224–229 (2015).
4. Padilla, J. & Vanderbilt, D. *Surf. Sci.* **418**, 64–70 (1998).

5. Bhalla, A. S. & Newnham, R. E. *Phys. Status Solidi A* **58**, K19–K24 (1980).
6. Gehring, P. M., Hirota, K., Majkrzak, C. F. & Shirane, G. *Phys. Rev. Lett.* **71**, 1087 (1993).
7. Domingo, N., Bagués, N., Santiso, J. & Catalan, G. *Phys. Rev. B* **91**, 094111 (2015).
8. Hirota, K., Hill, J. P., Shapiro, S. M., Shirane, G. & Fujii, Y. *Phys. Rev. B* **52**, 13195–13205 (1995).
9. Jiang, Q. & Zegenhagen, J. *Surf. Sci.* **367**, L42–L46 (1996).
10. Sai, N., Meyer, B. & Vanderbilt, D. *Phys. Rev. Lett.* **84**, 5636 (2000).
11. Yamada, H., Kawasaki, M., Ogawa, Y. & Tokura, Y. *Appl. Phys. Lett.* **81**, 4793 (2002).
12. Tagantsev, K. *Phys. Rev. B* **34**, 5883–5889 (1986).
13. Narvaez, J., Vasquez-Sancho, F. & Catalan, G. *Nature* **538**, 219–221 (2016).
14. Stengel, M. *Phys. Rev. B* **90**, 201112(R) (2014).
15. Stanley, H. E. *Rev. Mod. Phys.* **71**, S358 (1999).
16. Santander-Syro, A. F. *et al. Nature* **469**, 189–193 (2011).
17. Anderson, P. W. & Blount, E. I. *Phys. Rev. Lett.* **14**, 217–219 (1965).
18. Shi, Y. *et al. Nature Mater.* **12**, 1024–1027 (2013).
19. Benedek, N. A. & Biroli, T. *J. Mater. Chem. C* **4**, 4000–4015 (2016).

This article was published online on 18 November 2019.

Evolution

The balancing act of growth and expansion

Henry Mattingly & Thierry Emonet

Bacteria move along gradients of chemical attractants. Two studies find that, in nutrient-rich environments, bacteria can grow rapidly by following a non-nutritious attractant – but expanding too fast leaves them vulnerable. See p.658 & p.664

Bacteria can sense chemical attractants and use that information to navigate towards resources or away from harm – a process called chemotaxis. But why bacteria chase signals that often do not have much nutritional value has been a long-standing puzzle. Cremer *et al.*¹ show on page 658 that bacterial populations can use non-nutritious attractants as cues for rapidly expanding through nutrient-rich areas, ensuring that plentiful nutrients are available for their future growth. And on page 664, Liu *et al.*² build on this work to reveal an unanticipated rule of bacterial evolution: the safest way for a bacterial population to colonize a habitat is not necessarily to expand as fast as possible, because rapid expansion can leave the population vulnerable to invasion by competitors.

In the 1960s, the biochemist Julius Adler demonstrated that a group of cells consuming a chemical attractant can form a rapidly expanding wave that follows a moving concentration gradient that the cells create on their own³. That is, by consuming the attractant in their immediate vicinity, the cells create a gradient between their current location and the

surrounding regions in which the chemical has not yet been consumed. The cells then chase the higher concentration – rather like a horse chasing a carrot on a stick. The wave's

expansion speed is determined by how fast the travelling cells deplete the local attractant⁴.

Cremer *et al.* examined how a cell population's use of chemotaxis to expand (defined as the occupation of more space), as in Adler's experiments, affects its growth (the increase in cell numbers). The authors seeded small colonies of bacteria in a Petri dish, and measured population size over time as the cells grew and filled the available space. As Adler had observed, the colonies formed expanding waves, and some cells fell behind the wavefront, seeding the newly covered ground.

Importantly, when Cremer and colleagues added small amounts of a non-nutritious chemical attractant that was different from the nutrient on which the cells were growing, the population capitalized on chemotaxis to expand before the local nutrient had become depleted. This increased the number of cells that had access to nutrients at a given time and allowed the population to grow much faster than it did without the directional cue of the attractant (Fig. 1). This gain relied on a separation between chemotaxis and growth: the attractant served as a cue, rather than as a nutrient source⁵, to direct the cells towards unoccupied territory. When the attractant is the only nutrient, the population does not grow as fast; either the attractant is abundant, and the cells can't consume it fast enough for rapid expansion, or the attractant is limited, and expansion can be fast but the settlers behind the wavefront are starved and don't grow.

This work demonstrates that – in a nutrient-rich environment – the faster a single population expands, the faster it grows. But what happens when competitors (including spontaneously generated mutants in the population) expand into the same territory? Last year, we and our colleagues⁶ showed that bacteria with different chemotactic abilities

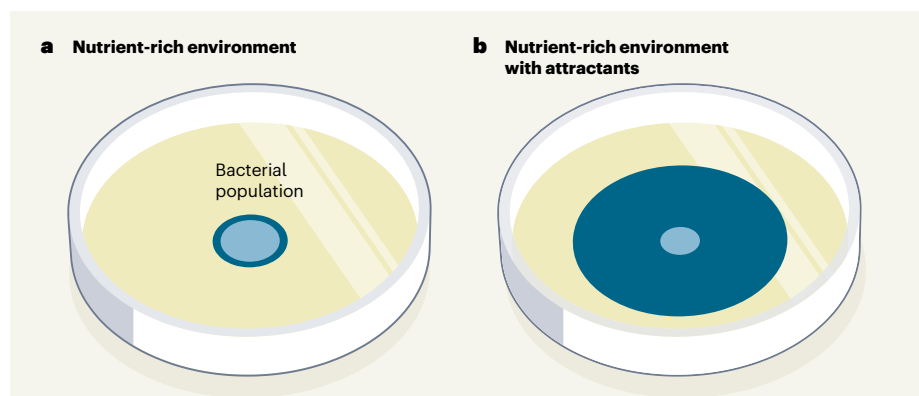


Figure 1 | How bacteria maximize growth in nutrient-rich environments. **a**, Populations of bacteria can spread out within a nutrient-rich habitat through cell division and random motion. But this approach causes most of the population to stay in a small location and deplete local nutrients – so, many cells starve (light blue) and only the outer edge grows (dark blue). **b**, Cremer *et al.*¹ grew bacteria in the same environment but included low levels of a non-nutritious attractant chemical (not shown). The cells chased the attractant through a process called chemotaxis, expanding rapidly across the dish before the local nutrients were depleted, so that most of the population had the nutrients needed to grow.

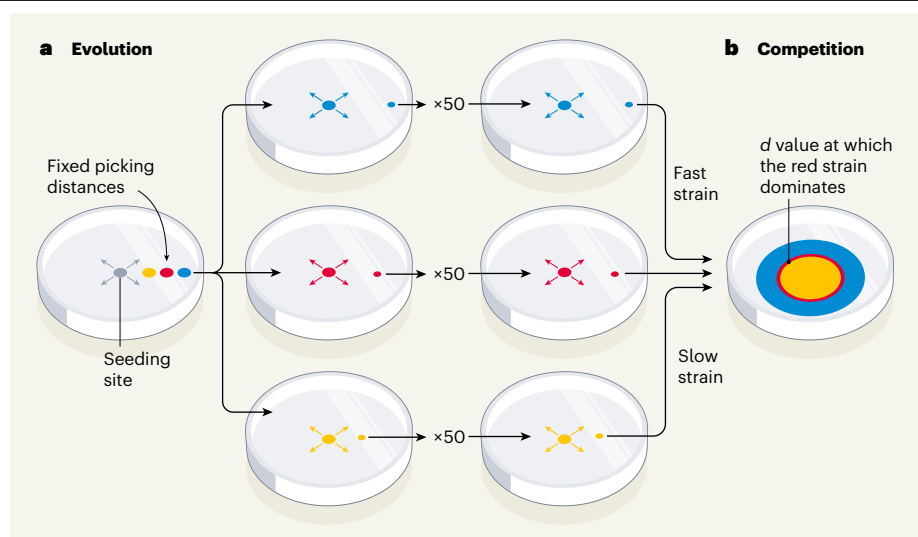


Figure 2 | An expansion strategy for protecting against competitors. Liu *et al.*² show that rapid expansion is not necessarily the best strategy if populations are competing for limited space. **a**, The authors seeded cells in the centre of a dish and let them expand across the dish (indicated by outward-pointing arrows). They then picked bacteria that had reached fixed distances from the seeding point (five distances were used, but only three are shown here, for simplicity). They reseeded cells picked from different distances in separate plates and repeated the process, picking cells from the same distance after the cells had filled the plate. Reseeding 50 times led to the evolution of strains that had expansion speeds that increased with picking distance. **b**, The team then seeded strains of different speeds together in a dish so that they would compete against one another. This revealed a simple rule for determining which strain will dominate at distance d from the seeding site – the one satisfying $d = u/\lambda$, where u is the strain's expansion rate when the population expands without competition and λ represents how quickly cells divide. Slow strains dominate close to the seeding site and fast strains dominate farther away, but at a particular value of d , the red strain cannot be outcompeted and is protected.

(but with the same genes) can travel together in the same expanding wave by spatially organizing themselves. High-performing cells travel at the front of the wave, where the attractant gradient is shallow; low-performing ones are found at the back, where the gradient is steeper because more of the attractant has been consumed. Steeper gradients are easier to navigate, so this spatial organization enables all cells to travel at about the same speed.

“The authors’ strains had evolved to fill a niche in which they were stable when facing invasion by competitors.”

However, cells at the back are more likely to fall behind the group and seed the covered ground. It has been unclear how this sorting mechanism affects the relative growth of multiple populations when they travel together.

Liu *et al.* addressed this question using an evolution experiment. As in Cremer and colleagues’ study, the authors seeded a population of bacteria in a Petri dish and allowed it to expand and fill the available space. Then the authors picked bacteria that had reached one of five fixed distances from the starting point and seeded them in a new dish in which

they expanded again (Fig. 2a). The researchers repeated the process 50 times, picking from the same distance each time.

Given that Cremer and colleagues found that faster expansion leads to greater growth, one might expect that Liu and colleagues’ protocol would select for strains that showed increasingly fast expansion, regardless of the picking distance. Instead, as the cycles progressed, strains picked at locations close to the initial seeding site evolved to expand more slowly than their ancestors, whereas those picked farther away evolved to expand more quickly.

Liu *et al.* then performed a competition assay in which they seeded evolved strains of different expansion speeds in the same dish. The authors found that the strains occupied different regions: slow strains deposited settlers behind the wave more rapidly and therefore dominated close to the seeding point, whereas fast strains deposited settlers more slowly and dominated far from it (Fig. 2b). Each strain’s fitness (quantified by its relative abundance) therefore depended on its distance from the starting point – a clue to the outcome of the evolution experiment.

Finally, through a combination of simulations and mathematical arguments, the researchers discovered a simple rule that predicts which strain is fittest at any given distance (d) from the seeding site. For

expansion speed u (when expanding without competition) and growth rate λ , the strain that satisfies $d = u/\lambda$ will dominate at d . As a strain’s expansion speed increases, fewer individuals of that strain fall behind to colonize the area, so it dominates at a greater distance (higher d), after other strains have lagged behind. By contrast, the higher a strain’s growth rate, the sooner it becomes the predominant strain in the expanding wavefront, and so the earlier it deposits settlers.

Taken together, these fascinating results show how populations of bacteria can balance rapid expansion and growth with ensuring that competitors cannot invade their territory. Liu and colleagues’ evolved strains were not the fittest in an absolute sense: they did not necessarily expand or grow the fastest when seeded in isolation. Rather, they had evolved to fill a niche in which they were stable when facing invasion by competitors.

The experimental systems developed in the two current studies are well suited for exploration of how the abilities of bacteria to shape and navigate their complex chemical environments affect ecological and evolutionary dynamics. These results should reach far beyond bacterial chemotaxis, and improve researchers’ understanding of the behaviour of growing populations across many fields.

Henry Mattingly and Thierry Emonet

are in the Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520-8103, USA.
e-mails: henry.mattingly@yale.edu;
thierry.emonet@yale.edu

1. Cremer, J. *et al.* *Nature* **575**, 658–663 (2019).
2. Liu, W., Cremer, J., Li, D., Hwa, T. & Liu, C. *Nature* **575**, 664–668 (2019).
3. Adler, J. *Science* **153**, 708–716 (1966).
4. Keller, E. F. & Segel, L. A. *J. Theor. Biol.* **30**, 235–248 (1971).
5. Yang, Y. *et al.* *Mol. Microbiol.* **96**, 1272–1282 (2015).
6. Fu, X. *et al.* *Nature Commun.* **9**, 2177 (2018).

This article was published online on 6 November 2019.

Quantum dots realize their potential

Alexander L. Efros

Scientists have engineered semiconducting nanocrystals called quantum dots that lack toxic heavy metals and are highly efficient light emitters. These nanostructures might be used in displays, solar cells and light-emitting diodes. **See p.634**

Tiny semiconductor crystals dubbed quantum dots (QDs) are one of the biggest nanotechnology success stories so far. Since their first synthesis^{1,2} in the 1980s, QDs have featured in a wide range of optoelectronic devices, and QDs suspended in solution have been used in many *in vivo* and *in vitro* imaging, labelling and sensing techniques. However, two technical problems need to be resolved before their potential can be fully realized. First, QDs based on cadmium must be replaced by ones that are highly efficient light emitters and that do not contain such toxic heavy metals. And second, QD phosphors (substances that exhibit luminescence) in televisions must be replaced by QD light-emitting diodes (LEDs), to reduce power consumption. On page 634, Won *et al.*³ report QDs that address both issues.

The absorption spectra of nanocrystal QDs depend on their size. This property was discovered independently for QDs in glass¹ and in aqueous solution² and was first described quantitatively^{4,5} in the early 1980s. For practical applications, such a feature should be converted into size-dependent photoluminescence. In this process, an electron in the valence energy band of a QD absorbs a photon and moves to the conduction energy band, leaving behind a hole (electron vacancy). The photoexcited electron and hole then recombine (merge), releasing a photon (Fig. 1a).

Photoluminescence was achieved initially by coating QD surfaces with organic molecules⁶ and later by using QDs that comprise a semiconductor core surrounded by a shell of a semiconductor that has a large band-gap⁷ – the energy difference between the valence and conduction bands. In the latter case, the offset in energy between the bands of the shell and those of the semiconductor core prevents electrons and holes from the core escaping to the external surface and enables intrinsic photoluminescence. In cadmium selenide QDs, the size-dependent absorption and photoluminescence spectra cover the entire range of visible wavelengths

from red to deep green⁸.

For QDs grown using current techniques, the photoluminescence quantum yield (the number of photons emitted by the QD divided by the number of photons absorbed) can be quite high. However, this high quantum yield is still not good enough for some applications. For instance, a quantum yield of less than 100% is associated with blinking⁹ – a phenomenon observed in single-QD experiments, in which the photoluminescence intensity varies under constant illumination. This blinking is linked to random processes in which QDs become charged and are subsequently neutralized.

An electron–hole pair that is photoexcited in a neutral QD can recombine only by emitting a photon – in other words, by photoluminescence. However, photoexcitation in a charged QD triggers another recombination process, which is known as non-radiative Auger recombination. In this process, the energy of the photoexcited electron–hole

pair is transferred to another electron or hole and a photon is not emitted (Fig. 1b). For commonly used QDs that comprise a core surrounded by a thin shell, the rate of Auger recombination is usually much higher than that of photoluminescence. As a result, the former process completely quenches the latter process in most charged QDs.

To achieve a photoluminescence quantum yield that is close to 100%, Auger recombination needs to be suppressed. One approach is to prevent the optically produced electrons and holes from escaping to the QD surface, to avoid charging of the QD. This can be realized, for example, by using QDs that have a thick shell⁹. In the case of QD-LEDs, QD neutrality can be controlled by ensuring that electron and hole conductivities are similar.

A complementary approach is to engineer the QDs to have a soft confinement potential – a potential-energy profile that gently increases at the QD surface and so reduces the rate of Auger recombination¹⁰. This potential can be produced by forming an alloy of the core and shell materials at the core–shell interface of QDs, or by using multi-shell QDs in which each subsequent shell has a larger bandgap than the preceding one. Successful efforts to engineer such QDs were reported last year¹¹. However, despite the outstanding optical properties of the cadmium selenide-based QD structures that were attained in that work, the photoluminescence quantum yield did not reach 100%.

Won and colleagues have developed an innovative method for synthesizing heavy-metal-free QDs that consist of a uniform indium phosphide core, a thick inner shell of zinc selenide and a thin outer shell of zinc sulfide (Fig. 1c). The technique involves two consecutive steps for the growth of the

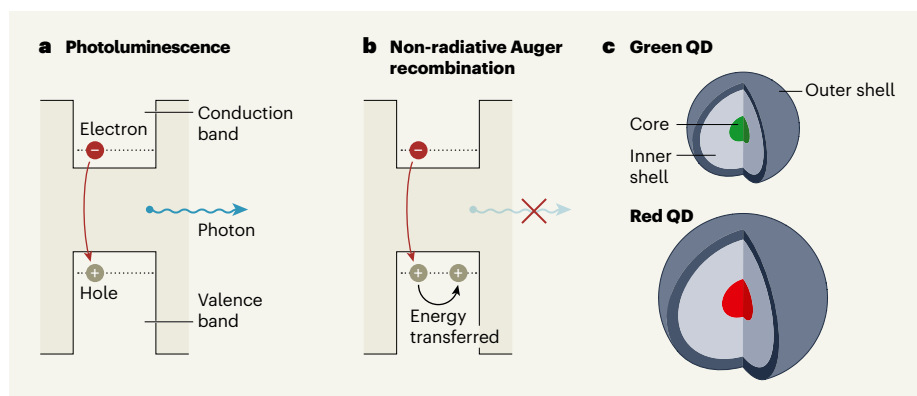


Figure 1 | Efficient light emission from quantum dots. **a**, Semiconducting nanocrystals known as quantum dots (QDs) can produce light through photoluminescence. In this process, there is a negatively charged electron in the conduction energy band of the QD and a positively charged hole (electron vacancy) in the valence energy band. The electron moves to the valence band and merges with the hole, releasing a photon. **b**, However, a process called non-radiative Auger recombination can instead occur. In this process, the energy generated by the electron–hole merger is transferred to another charge carrier, and a photon is not emitted. **c**, Won *et al.*³ demonstrate QDs that are highly efficient light emitters, because Auger recombination is suppressed. The QDs consist of a uniform indium phosphide core, a thick inner shell of zinc selenide and a thin outer shell of zinc sulfide. The colour of the light produced by the core depends on its size.

core: the addition of hydrofluoric acid to etch off the oxidized core surface during the growth of the initial zinc selenide shell, and high-temperature zinc selenide growth at 340 °C.

The resulting QDs have a highly symmetrical spherical shape, which is one of the conditions for realizing a soft confinement potential. Any cavity or sharp corner on the surface or at the core-shell interface would enhance the rate of non-radiative Auger recombination. Charged or deep-level defects would also lead to such enhancement. The authors found that the thick zinc selenide shell suppresses Auger recombination, suggesting that the interface is of extremely high quality and that there are no crystal defects called stacking faults in the zinc selenide shell. The intrinsic photoluminescence quantum yield of these QDs is 100%.

Won *et al.* used their QDs to make LED devices in which electrons and holes are

injected into the QDs instead of being photoexcited. To maintain QD neutrality during such injection, and to improve the transport of electrons and holes in the LEDs, the authors replaced long-chain ligand molecules at the QD surface with short-chain ones. The QD-LEDs achieve an external quantum efficiency (the number of photons that leave the LED divided by the number of charges injected into it) of 21.4%, which is the theoretical maximum. The improved injection and transport of charges reduces accumulated electrical resistance during operation, lowers power consumption and increases the lifetime of the LED device.

This work shows that the detailed understanding of the physical properties of QDs that has accumulated over more than 30 years should now allow us to engineer QDs for multiple and diverse applications. These could include televisions and displays, LEDs and solar cells.

Alexander L. Efros is at the Center for Computational Material Science, Naval Research Laboratory, Washington DC 20375, USA.

e-mail: efros@nrl.navy.mil

1. Ekimov, A. I. & Onushchenko, A. A. *JETP Lett.* **34**, 345–349 (1981).
2. Rossetti, R., Nakahara, S. & Brus, L. E. *J. Chem. Phys.* **79**, 1086–1088 (1983).
3. Won, Y.-H. *et al. Nature* **575**, 634–638 (2019).
4. Efros, A. L. & Efros, A. L. *Sov. Phys. Semicond.* **16**, 772–775 (1982).
5. Brus, L. E. *J. Chem. Phys.* **79**, 5566–5571 (1983).
6. Murray, C. B., Norris, D. J. & Bawendi, M. G. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
7. Hines, M. A. & Guyot-Sionnest, P. *J. Phys. Chem.* **100**, 468–471 (1996).
8. Efros, A. L. *et al. Phys. Rev. B* **54**, 4843–4856 (1996).
9. Efros, A. L. & Nesbitt, D. J. *Nature Nanotechnol.* **11**, 661–671 (2016).
10. Cragg, G. E. & Efros, A. L. *Nano Lett.* **10**, 313–317 (2010).
11. Lim, J., Park, Y.-S. & Klimov, V. I. *Nature Mater.* **17**, 42–49 (2018).

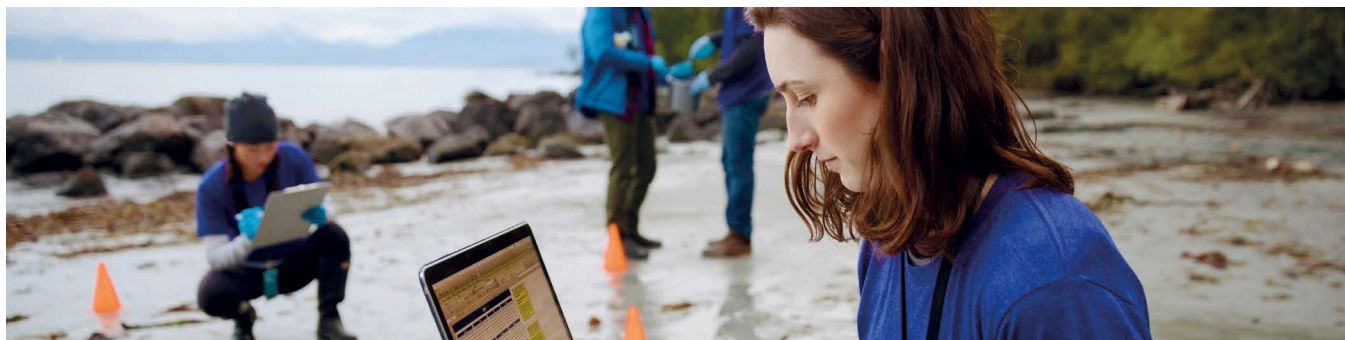
nature masterclasses

Online Course in Scientific Writing and Publishing

Delivered by Nature Research journal editors, researchers gain an unparalleled insight into how to publish.



Try a free sample of the course at masterclasses.nature.com



Bite-size design for busy researchers • Subscribe as a lab or institution

[W masterclasses.nature.com](https://masterclasses.nature.com)

[in](#) Follow us on LinkedIn

[f](#) Skills and Careers Forum for Researchers

A80768

Towards spike-based machine intelligence with neuromorphic computing

<https://doi.org/10.1038/s41586-019-1677-2>

Kaushik Roy^{1*}, Akhilesh Jaiswal¹ & Priyadarshini Panda¹

Received: 23 July 2018

Accepted: 9 July 2019

Published online: 27 November 2019

Guided by brain-like ‘spiking’ computational frameworks, neuromorphic computing—brain-inspired computing for machine intelligence—promises to realize artificial intelligence while reducing the energy requirements of computing platforms. This interdisciplinary field began with the implementation of silicon circuits for biological neural routines, but has evolved to encompass the hardware implementation of algorithms with spike-based encoding and event-driven representations. Here we provide an overview of the developments in neuromorphic computing for both algorithms and hardware and highlight the fundamentals of learning and hardware frameworks. We discuss the main challenges and the future prospects of neuromorphic computing, with emphasis on algorithm–hardware codesign.

Throughout history, the promise of creating technology with brain-like ability has been a source of innovation. Previously, scientists have contended that information transfer in the brain occurs through different channels and frequencies, as in a radio. Today, scientists argue that the brain is like a computer. With the development of neural networks, computers today have demonstrated extraordinary abilities in several cognition tasks—for example, the ability of AlphaGo to defeat human players at the strategic board game Go¹. Although this performance is truly impressive, a key question still remains: what is the computing cost involved in such activities?

The human brain performs impressive feats (for example, simultaneous recognition, reasoning, control and movement), with a power budget² of nearly 20 W. By contrast, a standard computer performing only recognition among 1,000 different kinds of objects³ expends about 250 W. Although the brain remains vastly unexplored, its remarkable capability may be attributed to three foundational observations from neuroscience: vast connectivity, structural and functional organizational hierarchy, and time-dependent neuronal and synaptic functionality^{4,5} (Fig. 1a). Neurons are the computational primitive elements of the brain that exchange or transfer information through discrete action potentials or ‘spikes’, and synapses are the storage elements underlying memory and learning. The human brain has a network of billions of neurons, interconnected through trillions of synapses. Spike-based temporal processing allows sparse and efficient information transfer in the brain. Studies have also revealed that the visual system of primates is organized as a hierarchical cascade of interconnected areas² that gradually transforms the representation of an object into a robust format, facilitating perceptive abilities.

Inspired by the brain’s hierarchical structure and neuro-synaptic framework, state-of-the-art artificial intelligence is, by and large, implemented using neural networks. In fact, modern deep-learning networks (DLNs) are essentially artefacts of hierarchy built by composing several layers or transformations that represent different latent features in the input⁶ (Fig. 1b). Such neural networks are fuelled by hardware computing systems that fundamentally rely on basic silicon transistors. Digital

logic in massive computing platforms comprises billions of transistors integrated on a single silicon die. Reminiscent of the hierarchical organization of the brain, various silicon-based computational aspects are arranged in a hierarchical fashion to allow efficient data exchange (see Fig. 1c).

Despite this superficial resemblance, there exists a sharp contrast between the computing principles of the brain and silicon-based computers. A few key differences include: (1) the segregation of computations (the processing unit) and storage (the memory unit) in computers contrasts with the co-located computing (neurons) and storage (synapses) mechanisms found in the brain; (2) the massive three-dimensional connectivity in the brain is currently beyond the reach of silicon technology, which is limited by two-dimensional connections and finite number of interconnecting metal layers and routing protocols; and (3) transistors are largely used as switches to construct deterministic Boolean (digital) circuits, in contrast to the spike-based event-driven computations in the brain that are inherently stochastic⁷. Nevertheless, silicon computing platforms (for example, graphics processing unit (GPU) cloud servers) have been one of the enabling factors in the current deep-learning revolution. However, a major bottleneck prohibiting the realization of ‘ubiquitous intelligence’ (spanning cloud-based servers to edge devices) is the large energy and throughput requirement. For example, running a deep network on an embedded smart-glass processor supported by a typical 2.1 W h battery would drain the battery completely within just 25 minutes (ref.⁸).

Guided by the brain, hardware systems that implement neuronal and synaptic computations through spike-driven communication may enable energy-efficient machine intelligence. Neuromorphic computing efforts (see Fig. 2) originated in the 1980s to mimic biological neuron and synapse functionality with transistors, quickly evolving to encompass the event-driven nature of computations (an artefact of discrete ‘spikes’). Eventually, in the early 2000s, such research efforts facilitated the emergence of large-scale neuromorphic chips. Today, the advantages and limitations of spike-driven computations (specifically, learning with ‘spikes’) are being actively explored by algorithm

¹Purdue University, West Lafayette, IN, USA. *e-mail: kaushik@purdue.edu

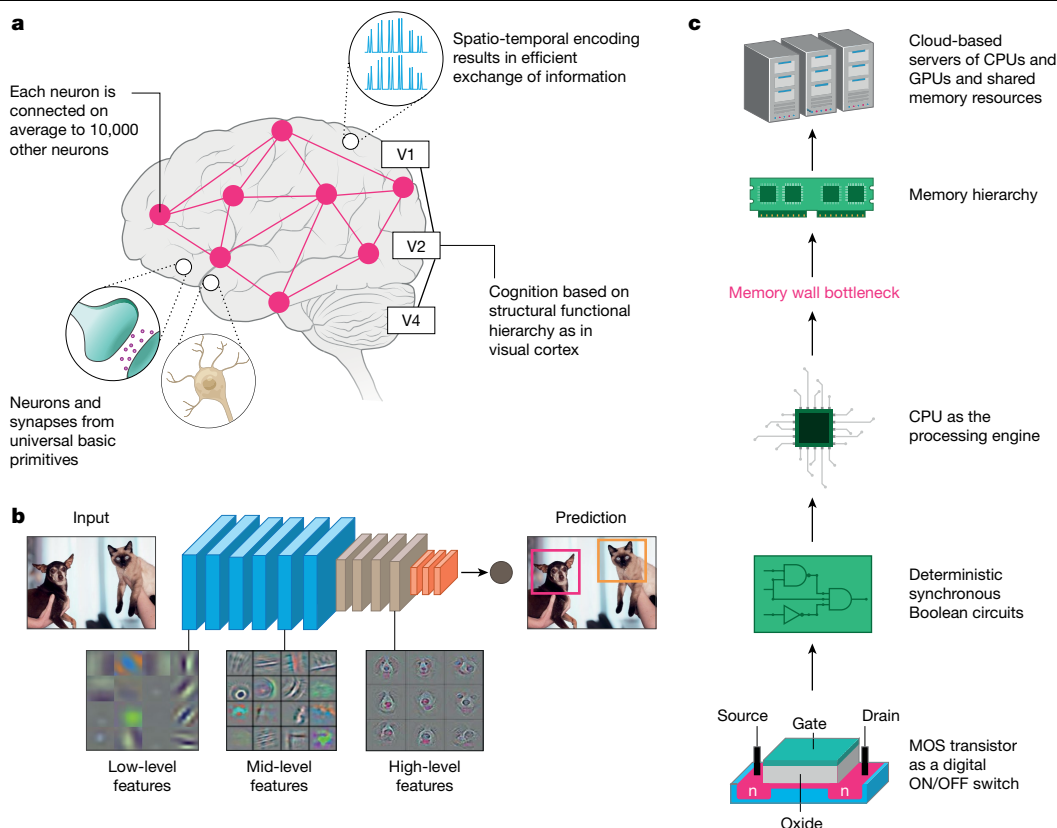


Fig. 1 | Key attributes of biological and silicon-based computing frameworks.

a, A schematic of the organizational principles of the brain. The intertwined network of neurons and synapses with temporal spike processing enables rapid and efficient flow of information between different areas.

b, A deep convolutional neural network performing objection detection on an image. These networks are multi-layered and use synaptic storage and neuronal nonlinearity that learn broad representations about the data. After training using backpropagation¹², the features learned at each layer show interesting patterns. The first layer learns general features such as edges and colour blobs. As we go deeper into the network, the learned features become

more specific, representing object parts (such as the eyes or nose of the dog) to full objects (such as the face of the dog). Such generic-to-specific transition is representative of the hierarchical arrangement of the visual cortex. **c**, A state-of-the-art silicon computing ecosystem. Broadly, the computing hierarchy is divided into processing units and memory storage. The physical separation of the processing unit and the memory hierarchy results in the well known ‘memory wall bottleneck’⁹⁴. Today’s deep neural networks are trained on powerful cloud servers, yielding incredible accuracy although incurring huge energy consumption.

designers to drive scalable, energy-efficient ‘spiking neural networks’ (SNNs). In this context, we can describe the field of neuromorphic computing as a synergistic effort that is equally weighted across both hardware and algorithmic domains to enable spike-based artificial intelligence. We first address the ‘intelligence’ (or algorithmic) aspects, including different learning mechanisms (unsupervised and supervised spike-based or gradient-descent schemes), while highlighting the need to exploit spatio-temporal event representations. A majority of this discussion focuses on applications for vision and related tasks, such as image recognition and detection. We then investigate the ‘computation’ (or hardware) aspects including analog computing, digital neuromorphic systems, beyond both von Neumann (the state-of-the-art architecture for digital computing systems) and silicon (representing the basic field-effect-transistor device that fuels today’s computing platforms) technology. Finally, we discuss the prospects of algorithm–hardware codesign wherein algorithmic resilience can be used to counter hardware vulnerability, thereby achieving the optimal trade-off between energy efficiency and accuracy.

Algorithmic outlook

Spiking neural networks

The seminal paper from Maass⁹ categorizes neural networks into three generations based on their underlying neuronal functionality. The first

generation, referred to as McCulloch–Pitt perceptrons, performs a thresholding operation resulting in a digital (1, 0) output¹⁰. The second generation—based on, for example, a sigmoid unit or a rectified linear unit¹¹ (ReLU)—adds continuous nonlinearity to the neuronal unit, which enables it to evaluate a continuous set of output values. This nonlinearity upgrade between the first- and second-generation networks had a key role in enabling the scaling of neural networks for complex applications and deeper implementations. Current DLNs, which have multiple hidden layers between input and output, are all based on such second-generation neurons. In fact, owing to their continuous neuronal functionality, these models support gradient-descent-based backpropagation learning¹²—the standard algorithm for training DLNs today. The third generation of networks use spiking neurons primarily of the ‘integrate-and-fire’ type¹³ that exchange information via spikes (Fig. 3).

The most important distinction between the second- and third-generation networks is in the nature of information processing. The former generation uses real-valued computation (say, the amplitude of the signal), whereas SNNs use the timing of the signals (or the spikes) to process information. Spikes are essentially binary events, either 0 or 1. As can be seen in Fig. 3a, a neuronal unit in an SNN is only active when it receives or emits spikes—it is therefore event-driven, which can contribute to energy efficiency over a given period of time. SNN units that do not experience any events remain idle. This is in contrast to DLNs, in which all units are active irrespective of the real-valued input or output

Algorithms

● Understanding the brain ● Enabling artificial intelligence

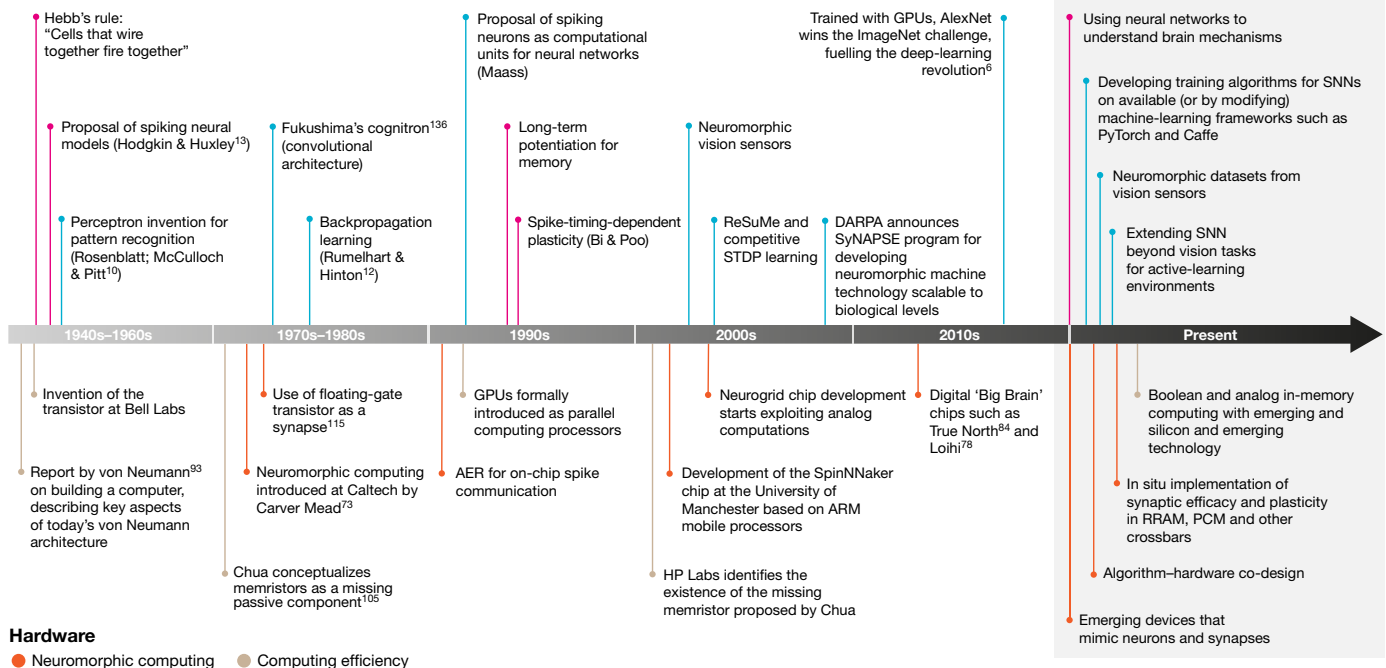


Fig. 2 | Timeline of major discoveries and advances in intelligent computing, from the 1940s to the present^{6, 10, 14, 73, 78, 84, 93, 105, 115, 136, 150, 151}.

For hardware, we have indicated discoveries from two perspectives—those motivated towards neuromorphic computing or that have enabled brain-like computations and 'intelligence' with hardware innovations; and those motivated towards computing efficiency, or that have enabled faster and more energy-efficient Boolean computations. From an algorithmic perspective, we have indicated the discoveries

as motivated towards understanding the brain, that is, driven by neuroscience and biological sciences; and motivated towards enabling artificial intelligence, that is, driven by engineering and applied sciences. Note that this is not a complete or comprehensive list of all discoveries. 'Current research' does not necessarily imply that such efforts have not been explored in the past; instead, we have emphasized key aspects of ongoing and promising research in the field.

values. Furthermore, the fact that the inputs in an SNN are either 1 or 0 reduces the mathematical dot-product operation, $\sum_i V_i \times w_i$ (detailed in Fig. 3a), to a less computationally intensive summation operation.

Different spiking neuron models, such as leaky integrate-and-fire (LIF) (Fig. 3b) and Hodgkin–Huxley¹³, have been proposed to describe the generation of spikes at different levels of bio-fidelity. Similarly, for synaptic plasticity, schemes such as Hebbian¹⁴ and non-Hebbian have been proposed¹⁵. Synaptic plasticity—the modulation of synaptic weights, which translates to learning in SNNs—relies on the relative timing of pre- and post-synaptic spikes (Fig. 3c). A suitable spiking neuron model with proper synaptic plasticity while exploiting event-based, data-driven updates (with event-based sensors^{16,17}) is a major goal among neuromorphic engineers, to enable computationally efficient intelligence applications such as recognition and inference, among others.

Exploiting event-based data with SNNs

We believe that the ultimate advantage of SNNs comes from their ability to fully exploit spatio-temporal event-based information. Today, we have reasonably mature neuromorphic sensors^{16,18} that can record dynamic changes in activity from an environment in real-time. Such dynamic sensory data can be combined with the temporal processing capability of SNNs to enable extremely low-power computing. In fact, using time as an additional input dimension, SNNs record valuable information in a sparse manner, compared with the frame-driven approaches that are traditionally used by DLNs (see Fig. 3). This can lead to efficient implementation of SNN frameworks, computing optical visual flow^{19,20} or stereo vision to achieve depth perception^{21,22}, that, in combination with spike-based-learning rules, can yield efficient training. Researchers in the robotics community have already demonstrated the benefit of

using event-based sensors for tracking and gesture recognition, among other applications^{19,21,22}. However, most of these applications use a DLN to perform recognition.

A major restriction in the use of SNNs with such sensors is the lack of appropriate training algorithms that can efficiently utilize the timing information of the spiking neurons. Practically, in terms of accuracy, SNNs are still behind their second-generation deep-learning counterparts in most learning tasks. It is evident that spiking neurons have a discontinuous functionality, and emit discrete spikes that are non-differentiable (see Fig. 3); hence they cannot use the gradient-descent backpropagation techniques that are fundamental to conventional neural network training.

Another restriction on SNNs is spike-based data availability. Although the ideal situation requires the input to SNNs to be spike trains with timing information, the performance of SNN training algorithms is evaluated on existing static-image datasets, for example CIFAR²³ or ImageNet²⁴, for recognition. Such static-frame-based data are then converted to spike trains using appropriate encoding techniques, such as rate coding or rank-order coding²⁵ (see Fig. 3d). Although encoding techniques enable us to evaluate the performance of SNNs on traditional benchmark datasets, we need to move beyond static-image classification tasks. The ultimate competence of SNNs should arise from their capability to process and perceive continuous input streams from the ever-changing real world, just as our brains do effortlessly. At present, we have neither good benchmark datasets nor the metrics to evaluate such real-world performance of SNNs. More research into gathering appropriate benchmark datasets, such as dynamic vision sensor data²⁶ or driving and navigation instances^{27,28}, is vital.

(Here we refer to the second-generation continuous neural networks as DLNs to differentiate them from spike-based computing. We note

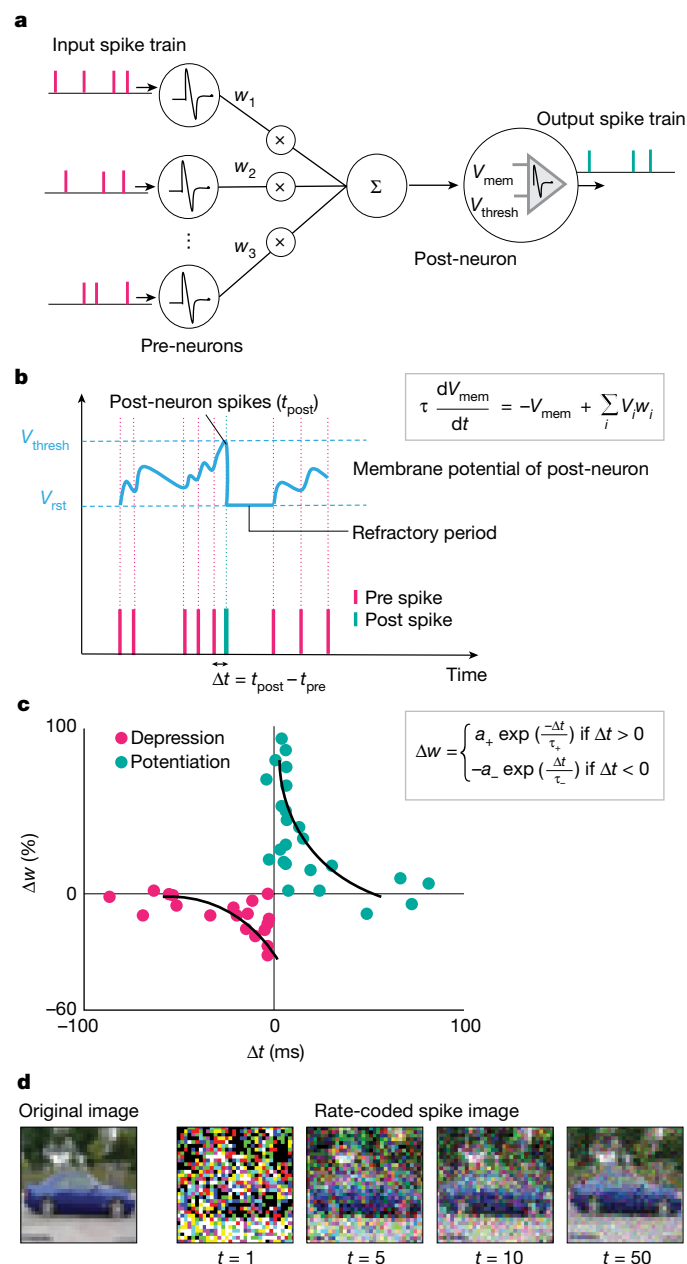


Fig. 3 | SNN computational models. **a**, A neural network, comprising a post-neuron driven by input pre-neurons. The pre-neuronal spikes, V_i are modulated by synaptic weights, w_i to produce a resultant current, $\sum_i V_i \times w_i$ (equivalent to a dot-product operation) at a given time. The resulting current affects the membrane potential of the post-neuron. **b**, The dynamics of LIF spiking neurons is shown. The membrane potential, V_{mem} integrates incoming spikes and leaks with time constant, τ in the absence of spikes. The post-neuron generates an outgoing spike whenever V_{mem} crosses a threshold, V_{thresh} . A refractory period ensues after spike generation, during which V_{mem} of the post-neuron is not affected. **c**, The spike-timing-dependent plasticity (STDP) formulation based on experimental data is shown, where a_+ , a_- , τ_+ and τ_- are learning rates and time-constants governing the weight change, Δw . The synaptic weights w_i are updated on the basis of the time difference ($\Delta t = t_{\text{post}} - t_{\text{pre}}$) between the pre-neuron and post-neuron spikes. **d**, An input image (static-frame data) is converted to a spike map over various time steps using rate coding. Each pixel generates a Poisson spike train with a firing rate proportional to the pixel intensity. When the spike maps are summed over several time steps (the spike map labelled $t = 5$ is a summation of maps from $t = 1$ to $t = 5$), they start to resemble the input. Hence, spike-based encoding preserves the integrity of the input image and also binarizes the data in the temporal domain. It is evident that LIF behaviour and random-input spike-generation bring stochasticity to the internal dynamics of an SNN. Note that rank-order coding can also be used to generate spike data²⁵.

that SNNs can also be implemented on a deep architecture with convolutional hierarchy while performing spiking neuronal functions.)

Learning in SNNs

Conversion-based approaches

The idea of a conversion-based approach is to obtain an SNN that yields the same input–output mapping for a given task as that of a DLN. Essentially, a trained DLN is converted to an SNN using weight rescaling and normalization methods to match the characteristics of a nonlinear continuous output neuron to that of the leak time constants, refractory period, membrane threshold and other functionalities of a spiking neuron^{29–34}. Such approaches have thus far been able to yield the most competitive accuracy on large-scale spiking networks in image classification (including on the ImageNet dataset). In conversion-based approaches, the advantage is that the burden of training in the temporal domain is removed. A DLN is trained on frame-based data using available frameworks such as Tensorflow³⁵ that offer training-related flexibility. Conversion requires parsing the trained DLN on event-based data (obtained by rate coding of the static-image dataset) and then applying simple transformations. However, such methods have inherent limitations. The output value of a nonlinear neuron—using, for example, a hyperbolic tangent (tanh) or a normalized exponential (softmax) function—can take both positive and negative values, whereas the rate of a spiking neuron can only be positive. Thus, negative values will always be discarded, leading to a decline in accuracy of the converted SNNs. Another problem with conversion is obtaining the optimal firing rate at each layer without any drastic performance loss. Recent works^{29–31} have proposed practical solutions to determine optimal firing rates, and additional constraints (such as noise or leaky ReLUs) are introduced during training of the DLN to better match the spiking neuron's firing rate³⁶. Today, conversion approaches yield state-of-the-art accuracy for image-recognition tasks that parallel the classification performance of DLNs. It is noteworthy that the inference time for SNNs that are converted from DLNs turns out to be very large (of the order of a few thousand time steps), leading to increased latency as well as degraded energy efficiency.

Spike-based approaches

In a spike-based approach, SNNs are trained using timing information and therefore offer the obvious advantages of sparsity and efficiency in overall spiking dynamics. Researchers have adopted two main directions³⁷: unsupervised (training without labelled data), and supervised (training with labelled data). Early works in supervised learning were ReSuMe³⁸ and the tempotron³⁹, which demonstrate simple spike-based learning in a single-layer SNN using a variant of spike-timing-dependent plasticity (STDP) to perform classification. Since then, research efforts have been directed towards integrating global backpropagation-like spike-based error gradient descent to enable supervised learning in multi-layer SNNs. Most works that rely on backpropagation estimate a differentiable approximate function for the spiking neuronal functionality so that gradient descent can be performed (Fig. 4a). SpikeProp⁴⁰ and related variants^{41,42} have derived a backpropagation rule for SNNs by fixing a target spike train at the output layer. Recent works^{43–46} perform stochastic gradient descent on real-valued membrane potentials with the goal that the correct output neuron will fire more spikes randomly (instead of having a precise target spike train). These methods have achieved state-of-the-art results for deep convolutional SNNs for small-scale image recognition tasks such as digit classification on the MNIST (Modified National Institute of Standards and Technology) handwritten digits database⁴⁷. However, supervised learning—although more computationally efficient—has not been able to outperform conversion-based approaches in terms of accuracy for large-scale tasks.

On the other hand, inspired from neuroscience and with hardware-efficiency as the prime goal, unsupervised SNN training using local

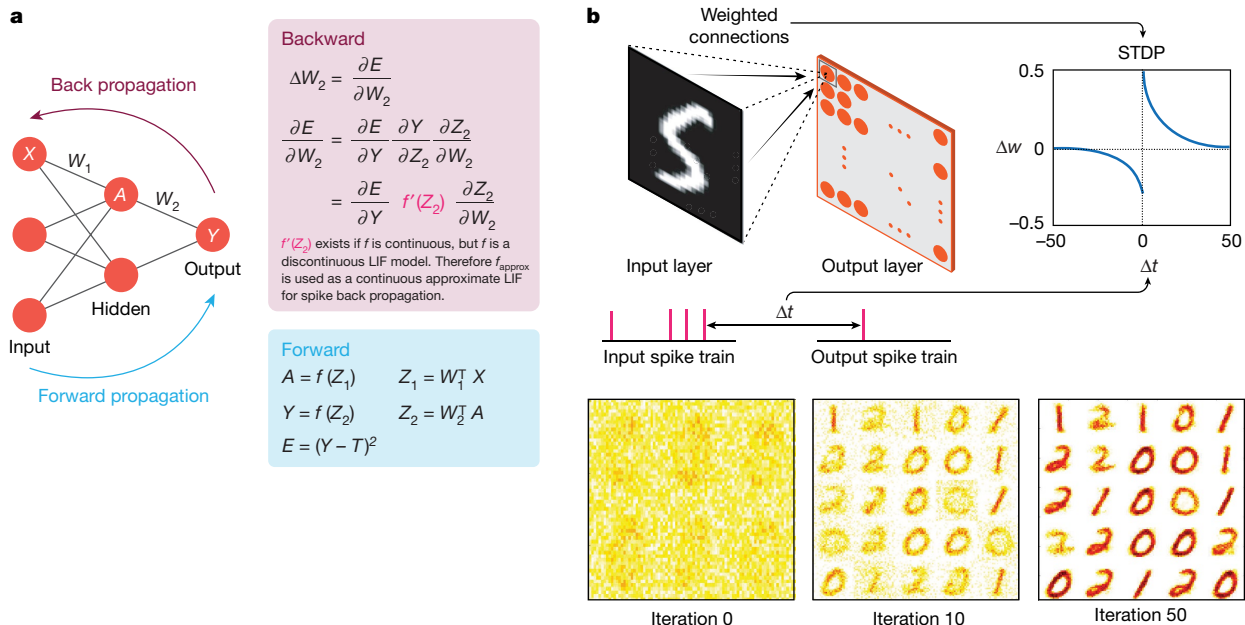


Fig. 4 | Global and local-learning principles in spiking networks.

a, Supervised global learning with known target labels, T for a classification task. Given a feedforward network, the network forward-propagates the input values, X through hidden layer units, A to yield an output, Y . The neuronal activation values, A at the hidden layer are calculated using the weighted summation of inputs—denoted $Z_1 = W_1^T X$ in matrix notation, combined with a nonlinear transformation, $f(Z_1)$. The outputs are calculated in a similar fashion. The derivative of the error, E with respect to the weights (W_1 , W_2) is then used to calculate the subsequent weight updates. Iteratively conducting the forward and backward propagation results in learning. The calculation of error derivatives requires f' , which necessitates that f is continuous. Consequently, the rules of spike-based backpropagation approximate the LIF function with

differentiable alternatives. The details of time-based information processing are not shown here. **b**, Local STDP unsupervised learning for digit classification. Given a two-layer topology with an input layer fully connected to all neurons in the output layer, the synaptic connections are learned through STDP. The weights are modulated on the basis of the difference in the spike timing of the input- and output-layer neurons. The weight value is increased (or decreased) if the input neuron fires before (or after) the output. With iterative training over multiple time steps, the weights—which were randomly initialized at the beginning—learn to encode a generic representation of a class of input patterns as shown (in this case, '0', '1' and '2'). Here, target labels are not required in order to perform recognition.

STDP-based learning rules⁴⁸ is also of great interest. With local learning (as we will see later in the hardware discussion), there are interesting opportunities to bring memory (synaptic storage) and computation (neuronal output) closer together. This architecture turns out to be more brain-like, as well as suitable for energy-efficient on-chip implementations. Diehl et al.⁴⁹ were one of the first to demonstrate completely unsupervised learning on an SNN, yielding comparable accuracy to deep learning on the MNIST database (Fig. 4b).

However, extending the local-learning approach to multiple layers for complex tasks is a challenge. As we go deeper into a network, the spiking probability (or the firing rate) of the neurons decreases, which we term 'vanishing forward-spike propagation'. To avoid this, most works^{46,48,50–53} train a multi-layer SNN (including convolutional SNNs) with local spike-based learning in a layer-wise fashion followed by global backpropagation learning, to perform classification. Such combined local–global approaches, although promising, are still behind conversion approaches in terms of classification accuracy. Further, recent works^{54,55} have shown proof-of-concept that the random projection of error signals through feedback connections in deep SNNs does enable improved learning. Such feedback-based learning methods need to be investigated further to estimate their efficacy on large-scale tasks.

Implications for learning in the binary regime

We can obtain extremely low-power and efficient computing with only binary (1/0) bit values rather than 16- or 32-bit floating point values that require additional memory. In fact, at the algorithmic level, learning in a probabilistic manner—wherein neurons spike randomly and weights have low-precision transitions—is being investigated to

obtain networks with few parameters and computation operations^{56–58}. Binary and ternary DLNs—in which the neuronal output and weights can take only the low-precision values -1 , 0 , and $+1$ —have been proposed, which yield good performance on large-scale classification tasks^{59,60}. SNNs already have a computational advantage as a result of binary spike-based processing. Furthermore, the stochasticity in neuronal dynamics of LIF neurons can improve the robustness of a network to external noise (for example, noisy input or noisy weight parameters from hardware)⁶¹. Then, it remains to be seen whether we can use this SNN temporal-processing architecture with appropriate learning methods, and compress weight training to a binary regime with minimal accuracy loss.

Other underexplored directions

Beyond vision tasks

So far, we have laid out approaches that have provided competitive results in, mostly, classification tasks. What about tasks beyond perception and inference on static images? SNNs offer an interesting opportunity for processing sequential data. However, there have been very few works³⁴ that have demonstrated the efficacy of an SNN in natural-language-processing tasks. What about reasoning and decision making with SNNs? Deep-learning researchers are heavily invested in reinforcement-learning algorithms that cause a model to learn by interacting with the environment in real time. Reinforcement learning with SNNs is very much underexplored^{62,63}. The current research efforts into SNNs—particularly in the area of training algorithms—shows that the grand challenge in SNNs is to match the

performance of deep learning. Although deep-learning serves as a good baseline for comparison, we believe that SNNs can create a niche for sensory-data processing, including in robotics and autonomous control.

Lifelong learning and learning with fewer data

Deep-learning models suffer from catastrophic forgetting when they undergo continual learning. For instance, when a network trained on task A is later exposed to task B, it forgets all about task A and remembers only task B. Establishing lifelong learning in a dynamically changing environment as humans do has garnered considerable attention from the research community. This is also a nascent direction in deep-learning research, but we need to think whether the additional temporal dimension of data processing in SNNs may help us to achieve continual learning⁶⁴. A similar direction worth exploring is one-shot learning. Learning with fewer data is the ultimate challenge and this is arguably one area where SNNs can achieve better results than deep learning. Unsupervised learning in SNNs can be combined with minimal supervision using only a fraction of the labelled training data to perform data-efficient training^{46,50,65}.

Forging links with neuroscience

We can take inspiration from neuroscience and apply those abstractions to learning rules in order to come up with efficient strategies. For instance, Masquelier et al.⁶⁵ employed STDP with temporal coding to mimic the visual-cortex pathway and found that such learning causes neurons to become feature selective—that is, different neurons learning different features—to different visual aspects of an image, resulting in a convolutional hierarchy of features. Similarly, incorporating dendritic learning⁶⁶ and structural plasticity⁶⁷ to improve spike-based learning by adding dendritic connections as an additional hyperparameter (a user-defined design parameter), offers interesting possibilities. A complementary body of work in the SNN domain is that of liquid state machines (LSMs)⁶⁸. LSMs use unstructured, randomly connected recurrent networks paired with a simple linear readout. Such frameworks with spiking dynamics have shown a surprising degree of success for a variety of sequential recognition tasks^{69–71}, but implementing them for complex and large-scale tasks remains an open problem.

Hardware outlook

From the above description of information processing and spike-based communication, a few characteristics of hardware systems that aim to form the underlying computational framework for SNNs can easily be hypothesized. Among these are the sparse-event-driven nature of the underlying hardware as a direct manifestation of the spike-based information exchange; the requirement for tightly intertwined computing and memory fabrics inspired by the ubiquitous presence of neurons and synapses throughout the biological brain; and the need to implement complex dynamical functions—for example, neuronal and synaptic dynamics using minimal circuit primitives.

The emergence of neuromorphic computing

In the 1980s, almost four decades after the invention of the transistor, Carver Mead envisioned “smarter” and “more-efficient” silicon computer fabrics based on certain aspects of biological neural systems^{72,73}. Although he suggested that his initial attempts to build such neuromorphic systems were “simple and stupid”⁷⁴, his work represented a new paradigm in hardware computing. Instead of focusing on Boolean computing based on basic AND and OR gates, Mead focused on analog distributed-computing circuits that mimicked neurons and synapses⁷⁴. He exploited the inherent device physics of the metal-oxide-silicon (MOS) transistor in the subthreshold regime—where current–voltage characteristics are exponential—to mimic

exponential neuronal dynamics⁷². Such device–circuit codesign is currently one of the most intriguing areas in neuromorphic computing, driven by novel emerging materials and associated devices.

The advent of parallel-processing GPUs

As opposed to CPUs (central processing units) that consist of one (or a few) complex computing core(s) integrated with on-chip memories, GPUs⁷⁵ consist of many simple computing cores that function in parallel, leading to high-throughput processing. GPUs were traditionally hardware accelerators for speeding up graphics applications. Of the many non-graphics applications that benefited from high-throughput computations of GPUs, deep learning is the most remarkable⁶. In fact, GPU servers are the go-to hardware platforms not only for running DLNs, but also for exploring inference and training for SNNs^{76,77}. While GPUs do provide an obvious advantage via their increased programming flexibility, they do not explicitly leverage the event-driven nature of spiking computations. In this regard, event-driven ‘Big Brain’ neuromorphic chips can yield the most energy-efficient solutions^{78,79}.

The ‘Big Brain’ chips

‘Big Brain’ chips⁸⁰ are distinguished by integrating millions of neurons and synapses that render spike-based computations^{78,81–86} (see Fig. 5a). Neurogrid⁸² and TrueNorth⁸⁴ are two model chips based on mixed-signal analog and digital circuits, respectively. TrueNorth uses digital circuits because analog circuits tend to accumulate errors easily, and are much more susceptible to process-induced variations in chip fabrication. Neurogrid was designed to assist computational neuroscience in emulating brain activity, with complex neuronal mechanisms such as opening and closing of various ion channels and the characteristic behaviour of biological synapses^{82,87}. By contrast, TrueNorth originated as a neuromorphic chip geared towards solving commercially important tasks such as recognition and classification using SNNs, and is based on simplified neural and synaptic primitives.

Taking the example of TrueNorth, two features that span different implementations of neuromorphic chips^{78,88} can be highlighted as follows.

Asynchronous address event representation. First, asynchronous address event representation (AER; Fig. 5b); this differs from conventional chip design, in which all computations are performed in parts with reference to a global clock. Because SNNs are sparse and computation is only required when a spike (or an event) is generated, asynchronous event-driven computation is much more suitable. In fact, enabling event-driven computations based on spikes has historically been one of the most attractive aspects of spike-based computations^{89,90}.

Network-on-chip. Second, networks-on-chip (NOCs) are used for spike communication. NOCs are networks of on-chip routers that receive and transmit packets of digital information through a time-multiplexed shared bus. The use of NOCs for large-scale chips is imperative, because connectivity in a typical silicon fabrication process is largely two-dimensional, with limited flexibility in the third dimension. We note that, despite the use of NOCs, on-chip connectivity still cannot rival the three-dimensional connectivity found in the brain. TrueNorth—and subsequent large-scale digital neuromorphic chips like Loihi⁷⁸—have demonstrated energy efficiency for SNN-based applications, taking us a step closer towards bio-fidelic implementations. However, limited connectivity, constrained bus bandwidth for NOCs and the all-digital approach remain key areas that require further investigation.

Beyond-von-Neumann computing

The sustained dimensional scaling of transistors—referred to as Moore’s law⁹¹—has driven the ever-increasing computing power of CPUs and GPUs as well as the ‘Big Brain’ chips. In recent years, this increase has

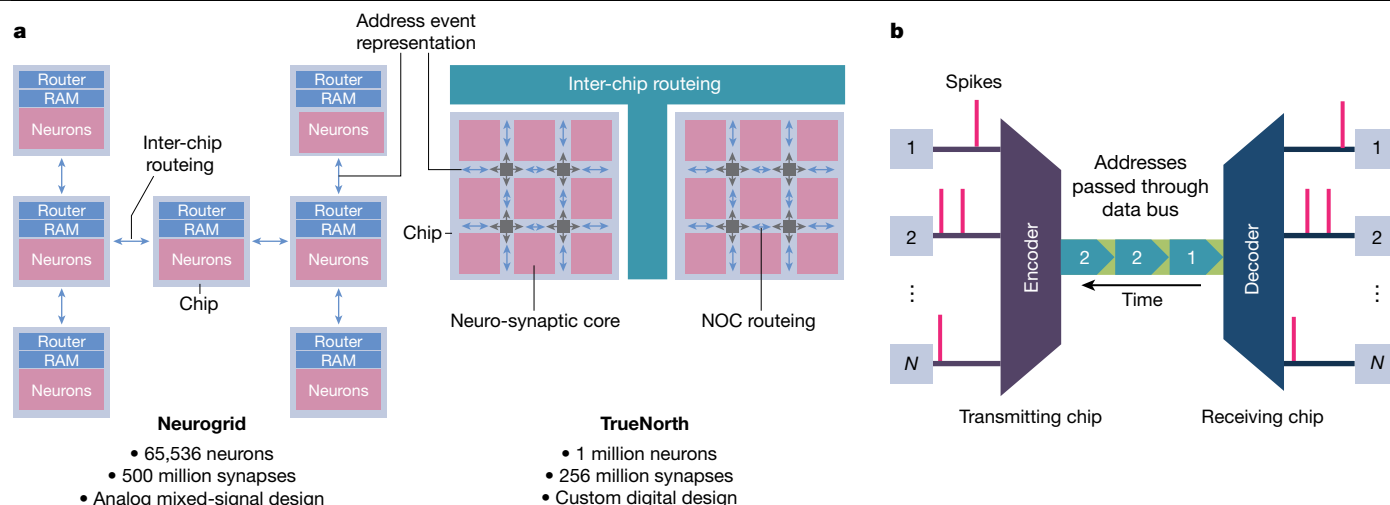


Fig. 5 | Some representative 'Big Brain' chips and AER methods. a, Among many works^{78,81–84} aimed at building large-scale neuromorphic chips, we highlight two representative systems—Neurogrid and TrueNorth. Neurogrid hosts more than 65,000 neurons and 500 million synapses, and TrueNorth has 1 million neurons and 256 million synapses. Neurogrid and TrueNorth use tree and mesh routing topology, respectively. Neurogrid uses an analog mixed-signal design and TrueNorth relies on digital primitives. In general, digital neuromorphic systems such as TrueNorth represent the membrane potential of a neuron as an n -bit binary word. Neuronal dynamics such as LIF behaviour are implemented by appropriately incrementing or decrementing the n -bit word. By contrast, analog systems represent the membrane potential as a charge stored on a capacitor. Current sources feeding into and sinking through

the capacitor node mimic the desired neuronal dynamics. Despite circuit differences, in general both analog and digital systems use event-driven AER for spike communication. Event-driven communication is one of the key enablers that allows integration of such large-scale systems, while simultaneously achieving low power dissipation. **b,** The basic AER communication system. Whenever an event (a spike) is generated on the transmitter side, the corresponding address is sent over the data bus to the receiver. The receiver decodes the incoming addresses and reconstructs the sequence of the spikes on the receiver side. Thus, each spike is explicitly encoded by its location (its address) and implicitly encoded by the time that its address is sent to the data bus.

slowed down as silicon-based transistors approach their physical limit⁹². To keep pace with soaring demand for computing power, researchers have recently begun exploring a two-pronged approach to enable both 'beyond von Neumann' and 'beyond silicon' computing models. A key shortcoming of the von Neumann model⁹³ is the clear demarcation of a processing unit physically separated from a storage unit, connected through a bus for data transfer (see Fig. 1c). The frequent movement of data between the faster processing unit and the slower memory unit through this bandwidth-constrained bus leads to the well-known 'memory wall bottleneck' that limits computing throughput and energy efficiency⁹⁴.

One of the most promising approaches in mitigating the effect of the memory wall bottleneck is to enable 'near-memory' and 'in-memory' computing^{95,96}. Near-memory computing enables co-location of memory and computing by embedding a dedicated processing engine in close proximity to the memory unit. In fact, the 'distributed computing architecture' of various 'Big Brain chips' (refer to Fig. 5) with closely placed neurons and synaptic arrays are representative of near-memory processing. By contrast, in-memory computing embeds certain aspects of computational operations within the memory array by enabling computation in the memory bit-cells or the peripheral circuits (see Fig. 6 for an example).

Non-volatile technologies

Non-volatile technologies^{97–103} are usually compared to biological synapses. In fact, they exhibit two of the most important characteristics of biological synapses: synaptic efficacy and synaptic plasticity. Synaptic plasticity is the ability to modulate the weights of the synapses based on a particular learning rule. Synaptic efficacy refers to the phenomenon of generating an output based on incoming spikes. In its simplest form, this means that incoming spikes are multiplied by the stored weights of synapses, which is usually represented as programmable, analog, non-volatile resistance. The multiplied signals are summed from all the

pre-neurons (neurons in a particular layer that receive input spikes) and applied as the input signal to the post-neuron (neurons in a particular layer that generate output spikes) (see Fig. 3). Figure 6 illustrates how in situ synaptic efficacy and synaptic plasticity can be accomplished using emerging non-volatile memristive technologies, arranged in a crossbar fashion^{103,104}. Additionally, such crossbars can be connected in an event-driven manner using NOCs to build dense, large-scale neuromorphic processors featuring in situ in-memory computations.

Various works based on memristive technologies^{105,106} such as resistive random-access memory (RRAM)¹⁰⁷, phase-change memory (PCM)¹⁰⁸ and spin-transfer torque magnetic random-access memory (STT-MRAM)¹⁰⁹ have been explored for both in situ dot-product computations and synaptic learning based on STDP rules. RRAMs (oxide-based and conductive-bridge-based¹⁰⁷) are electric-field-driven devices that rely on filament formation to achieve analog programmable resistance. RRAMs are prone to device-to-device and cycle-to-cycle variations^{110,111}, which is currently the major technical roadblock. PCMs comprise a chalcogenide material sandwiched between two electrodes that can switch its internal state between amorphous (high resistance) and crystalline (low resistance). PCM devices have comparable programming voltages and write speed to RRAMs although they suffer from high write-current and resistance drift over time¹⁰⁸. Spintronic devices consist of two magnets separated by a spacer; they exhibit two resistive states depending on whether the magnetization of the two layers is in the parallel or anti-parallel direction. Spin devices exhibit almost unlimited endurance, lower write energy and faster reversal compared to RRAMs and PCMs¹⁰⁹. However, the ratio of the two extreme resistive states (ON and OFF) is much smaller in spin devices than in PCMs and RRAMs.

Another class of non-volatile devices that allows tunable non-volatile resistance is a floating-gate transistor; such devices are being actively explored for synaptic storage^{112–114}. In fact, floating-gate devices were the first to be proposed as non-volatile synaptic storage^{115,116}. Because of their

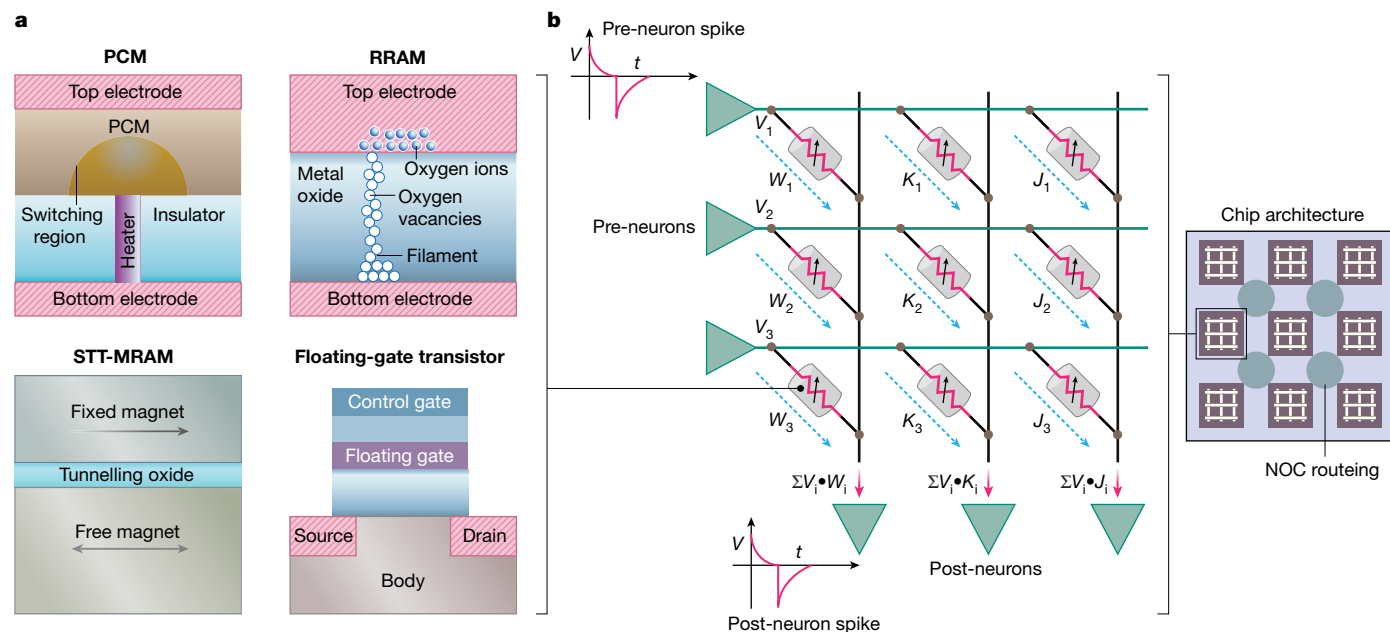


Fig. 6 | The use of non-volatile memory devices as synaptic storage.

a, Schematics of various non-volatile technologies: PCM, RRAM, STT-MRAM and floating-gate transistor. Such non-volatile devices have been used as synaptic storage and for in situ neuro-synaptic computations^{56,112–114,135,139–144}, and as in-memory accelerators for a wide range of generic non-neuromorphic applications^{128,145–149}. **b**, The implementation of synaptic efficacy and plasticity using memristive technologies. We show an array of memristors connected in a crossbar fashion. An incoming spike on the horizontal lines (green) results in a current that is proportional to the conductance of the memristive element in accordance with Ohm's law. Currents through multiple spiking pre-neurons are summed along vertical lines (black), a consequence of Kirchhoff's current law. This results in the in-memory dot-product operation that represents synaptic

efficacy. Synaptic plasticity is generally implemented in situ by appropriately applying a voltage pulse whenever the pre- and post-neurons spike on the horizontal and vertical lines, respectively, in accordance with a specific learning rule (as in STDP). The resistance values of the constituent memristors are programmed on the basis of the resulting voltage difference across the respective horizontal and vertical lines. The shape and timing of the voltage pulses to be applied for programming are chosen depending on the specific device technology. Note that floating-gate transistors, because they are three-terminal devices, require additional horizontal and/or vertical lines to enable crossbar functionality¹¹⁵. The figure also shows memristive arrays connected in a tiled fashion with NOCs that enable high-throughput in situ computations⁹⁷.

compatibility with MOS fabrication process, they are more mature than other emerging device technologies. However, the major limitation with floating-gate devices is their reduced endurance and high programming voltage in comparison to all other non-volatile technologies.

Although in situ computing and synaptic learning present attractive prospects for large-scale beyond-von-Neumann distributed computing, many challenges are yet to be overcome. Given device-to-device, cycle-to-cycle and process-induced variations, the approximate nature of computation is prone to errors that degrade overall computing efficiency as well as the accuracy of end applications. Further, the robustness of crossbar operation is affected by the presence of current sneak paths, line resistances, the source resistance of driving circuits and sensing resistance^{117,118}. Non-idealities of the selector device (either a transistor or a two-terminal nonlinear device), the requirement to have analog-digital converters and limited bit precision also add to the overall complexity of designing robust computing using non-traditional synaptic devices. Additionally, writing into non-volatile devices is usually energy intensive. Furthermore, the inherent stochastic nature of such devices can result in unreliable write operations that necessitate expensive and iterative write-verify schemes¹¹⁹.

Silicon (in-memory) computing

Apart from non-volatile technologies, various proposals for in-memory computing using standard silicon memories including static and dynamic random-access memories are under extensive investigation. Most of these works are focused on embedding Boolean bit-wise vector computations inside the memory arrays^{120–122}. Additionally, mixed-signal analog in-memory computing operations and binary convolutions have recently been demonstrated^{123,124}. In fact, in-memory

computing in various forms is currently being explored for almost all the major memory technologies, including static¹²⁵ and dynamic silicon memories¹²⁶, RRAMs¹²⁷, PCMs¹²⁸ and STT-MRAMs¹²⁹. Although most of these works have focused on generic computing applications like encryption and DLNs, they can easily find application in SNNs.

Algorithm-hardware codesign

Mixed-signal analog computing

Analog computing is highly susceptible to process-induced variations and noise, and is largely limited both in terms of area and energy consumption by the complexity and precision of analog and digital converters. Employing on-chip learning with tightly coupled analog computing frameworks will enable such systems to intrinsically adapt to process-induced variations, thereby mitigating their effect on accuracy. Localized learning with an emphasis on on-chip and on-device learning solutions has been investigated in the past^{130,131} and also in more recent bio-plausible algorithmic works⁵⁴. In essence, whether in the form of localized learning or in the use of paradigms like dendritic learning, we are of the opinion that a class of better error-resilient localized-learning algorithms—even at the cost of additional learning parameters—will be key in moving forward with analog neuromorphic computing. Additionally, the resilience of on-chip learning can be used to develop low-cost approximate analog-digital converters, without reducing the accuracy of a targeted application.

Memristive dot products

As a specific example of analog computing, memristive dot products are a promising approach towards enabling in situ neuromorphic

computing. Unfortunately, the resulting currents in memristive arrays representing the dot products have both spatial and data dependence, making crossbar circuit analysis a non-trivial, complex problem. Few works have studied the effect of crossbar non-idealities^{117,132,133} and explored training approaches to mitigate the effect of dot-product inaccuracies^{118,134}. Note that most of these works are focused on DLNs as opposed to SNNs. However, it is reasonable to assume that the basic device and circuit insights developed in these works are relevant for SNN implementations as well. Existing works require detailed device-circuit simulation runs that must be tightly coupled with training algorithms to diminish the accuracy loss. We believe an abstracted version of crossbar array models based on state-of-the-art devices, along with efforts to establish theoretical bounds in dot-product inaccuracies, are of immediate interest. This will enable an algorithm designer to explore new training algorithms while accounting for the hardware inconsistencies without time-consuming and iterative device-circuit-algorithm simulations.

Stochasticity

Stochastic SNNs are of substantial interest owing to the availability of emerging devices that are inherently stochastic^{135,136}. Most of the recent works on the implementation of stochastic binary SNNs have focused on small-scale tasks such as MNIST digit recognition⁵⁶. The common theme across such works is using stochastic STDP-like local learning rules to generate weight updates. We think that the temporal dimension in STDP learning provides additional bandwidth for weight updates to head in the right direction (towards achieving overall accuracy), even when constrained to the binary regime. The combination of such binary local-learning schemes with gradient-descent-based learning rules for large-scale tasks, while leveraging the stochasticity in hardware, provides interesting opportunities for energy-efficient neuromorphic systems.

Hybrid design approaches

We believe that hardware solutions based on hybrid approaches—that is, combining the advantages of various techniques on a single platform—is another important area that requires intensive investigation. Such approaches can be found in recent literature¹³⁷, where low-precision memristors are used in combination with a high-precision digital processor. There are many possible variants of such hybrid approaches, including significance-driven segregation of computational data, mixed-precision computations¹³⁷, reconfiguring conventional silicon memories as on-demand in-memory approximate accelerators¹²⁵, locally synchronous and globally asynchronous designs¹³⁸, locally analog and globally digital systems; wherein both emerging and silicon-based technologies can be used in unison to achieve improved accuracy and energy efficiency. Furthermore, such hybrid hardware can be used in tandem with hybrid spike-based learning approaches, such as locally unsupervised learning followed by globally supervised backpropagation⁵³. We believe that such combined local-global learning schemes can be leveraged to reduce hardware complexity, while also minimizing performance degradation for end applications.

Conclusion

Today, enabling ‘intelligence’ in almost all of the technology around us has become a central theme of research spanning various disciplines. In that regard, this Perspective sets out the case for neuromorphic computing as an energy-efficient way to enable machine intelligence through synergistic advancements in both hardware (computing) and algorithms (intelligence). We began by discussing the algorithmic implications of using a spiking neural paradigm, which uses event-driven computations, in contrast to real-valued computing in conventional deep-learning paradigms. We have described the advantages and limitations for realizing learning rules (such as spike-based

gradient-descent learning, unsupervised STDP and related conversion approaches from deep learning to the spiking domain) for standard classification tasks. Future algorithmic research should exploit the sparse and temporal dynamics of spike-based information processing, together with complementary neuromorphic datasets that can result in real-time recognition; and hardware development should focus on event-driven computations, co-location of memory and computational units, and mimicking dynamical neuro-synapse functionality. Of special interest are emerging non-volatile technologies enabling in situ mixed-signal analog computing. We have also discussed prospects for cross-layer optimization that enables algorithm-hardware code-sign—for example, exploiting algorithmic resilience (as in local learning) and hardware feasibility (as in ease of implementing stochastic primitives). Finally, the promise of spike-based energy-efficient and intelligent systems built with traditional and emerging devices is in sync with the current interest in enabling ubiquitous intelligence. Now is the time for the interchange of ideas, with multidisciplinary efforts spanning devices, circuits, architecture and algorithms to synthesize a truly energy-efficient and intelligent machine.

1. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
2. Cox, D. D. & Dean, T. Neural networks and neuroscience-inspired computer vision. *Curr. Biol.* **24**, R921–R929 (2014).
3. Milakov, M. Deep Learning With GPUs. <https://www.nvidia.co.uk/docs/IO/147844/Deep-Learning-With-GPUs-MaximMilakov-NVIDIA.pdf> (Nvidia, 2014).
4. Bullmore, E. & Sporns, O. The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336–349 (2012).
5. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* Vol. 28 (eds Pereira, F. et al.) 1097–1105 (Neural Information Processing Systems Foundation, 2012).
- This work—using deep convolutional networks—was the first to win the ImageNet challenge, fuelling the subsequent deep-learning revolution.**
7. Deco, G., Rolls, E. T. & Romo, R. Stochastic dynamics as a principle of brain function. *Prog. Neurobiol.* **88**, 1–16 (2009).
8. Venkataramani, S., Roy, K. & Raghunathan, A. Efficient embedded learning for IoT devices. In *21st Asia and South Pacific Design Automation Conf.* 308–311 (IEEE, 2016).
9. Maass, W. Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* **10**, 1659–1671 (1997).
- This paper was one of the first works to provide a rigorous mathematical analysis of the computational power of spiking neurons, categorizing them as the third generation of neural networks (after perceptron and sigmoidal neurons).**
10. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
11. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th Int. Conf. on Machine Learning* (eds Fürnkranz, J. & Joachims, T.) 807–814 (IMLS, 2010).
12. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
- This seminal work proposed gradient-descent-based backpropagation as a learning method for neural networks.**
13. Izhikevich, E. M. Simple model of spiking neurons. *IEEE Trans. Neural Netw.* **14**, 1569–1572 (2003).
14. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* (Wiley, 1949).
15. Abbott, L. F. & Nelson, S. B. Synaptic plasticity: taming the beast. *Nat. Neurosci.* **3**, 1178–1183 (2000).
16. Liu, S.-C. & Delbruck, T. Neuromorphic sensory systems. *Curr. Opin. Neurobiol.* **20**, 288–295 (2010).
17. Lichtsteiner, P., Posch, C. & Delbruck, T. A. 128×128 120 db 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits* **43**, 566–576 (2008).
18. Vanarse, A., Osseiran, A. & Rassau, A. A review of current neuromorphic approaches for vision, auditory, and olfactory sensors. *Front. Neurosci.* **10**, 115 (2016).
19. Benosman, R., Ieng, S.-H., Clercq, C., Bartolozzi, C. & Srinivasan, M. Asynchronous frameless event-based optical flow. *Neural Netw.* **27**, 32–37 (2012).
20. Wongsuphasawat, K. & Gotz, D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Trans. Vis. Comput. Graph.* **18**, 2659–2668 (2012).
21. Rogister, P., Benosman, R., Ieng, S.-H., Lichtsteiner, P. & Delbruck, T. Asynchronous event-based binocular stereo matching. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 347–353 (2012).
22. Osswald, M., Ieng, S.-H., Benosman, R. & Indiveri, G. A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems. *Sci. Rep.* **7**, 40703 (2017).
23. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. Preprint at <http://arxiv.org/abs/1207.0580> (2012).

24. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
25. Rullen, R. V. & Thorpe, S. J. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Comput.* **13**, 1255–1283 (2001).
26. Hu, Y., Liu, H., Pfeiffer, M. & Delbruck, T. DVS benchmark datasets for object tracking, action recognition, and object recognition. *Front. Neurosci.* **10**, 405 (2016).
27. Geiger, A., Lenz, P., Stiller, C. & Urtasun, R. Vision meets robotics: the KITTI dataset. *Int. J. Robot. Res.* **32**, 1231–1237 (2013).
28. Barranco, F., Fermüller, C., Aloimonos, Y. & Delbruck, T. A dataset for visual navigation with neuromorphic methods. *Front. Neurosci.* **10**, 49 (2016).
29. Sengupta, A., Ye, Y., Wang, R., Liu, C. & Roy, K. Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **13**, 95 (2019).
- This paper was the first to demonstrate the competitive performance of a conversion-based spiking neural network on ImageNet data for deep neural architectures.**
30. Cao, Y., Chen, Y. & Khosla, D. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comput. Vis.* **113**, 54–66 (2015).
31. Diehl, P. U. et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *Int. Joint Conf. on Neural Networks* 2933–2341 (IEEE, 2015).
32. Pérez-Carrasco, J. A. et al. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing—application to feedforward ConvNets. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 2706–2719 (2013).
33. Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M. & Liu, S.-C. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Front. Neurosci.* **11**, 682 (2017).
34. Diehl, P. U., Zarella, G., Cassidy, A. S., Pedroni, B. U. & Neftci, E. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *Int. Conf. on Rebooting Computing* 20 (IEEE, 2016).
35. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. In *12th USENIX Symp. Operating Systems Design and Implementation* 265–283 (2016).
36. Hunsberger, E. & Eliasmith, C. Spiking deep networks with LIF neurons. Preprint at <http://arxiv.org/abs/1510.08829> (2015).
37. Pfeiffer, M. & Pfeil, T. Deep learning with spiking neurons: opportunities and challenges. *Front. Neurosci.* **12**, 774 (2018).
38. Ponulak, F. & Kasiński, A. Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike shifting. *Neural Comput.* **22**, 467–510 (2010).
39. Güting, R. & Sompolinsky, H. The tempotron: a neuron that learns spike-timing-based decisions. *Nat. Neurosci.* **9**, 420–428 (2006).
40. Bohte, S. M., Kok, J. N. & La Poutré, H. Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing* **48**, 17–37 (2002).
41. Ghosh-Dastidar, S. & Adeli, H. A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection. *Neural Netw.* **22**, 1419–1431 (2009).
42. Anwani, N. & Rajendran, B. NormAD: normalized approximate descent-based supervised learning rule for spiking neurons. In *Int. Joint Conf. on Neural Networks* 2361–2368 (IEEE, 2015).
43. Lee, J. H., Delbruck, T. & Pfeiffer, M. Training deep spiking neural networks using backpropagation. *Front. Neurosci.* **10**, 508 (2016).
44. Orchard, G. et al. HFIRST: a temporal approach to object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 2028–2040 (2015).
45. Mostafa, H. Supervised learning based on temporal coding in spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 3227–3235 (2018).
46. Panda, P. & Roy, K. Unsupervised regenerative learning of hierarchical features in spiking deep networks for object recognition. In *Int. Joint Conf. on Neural Networks* 299–306 (IEEE, 2016).
47. LeCun, Y., Cortes, C. & Burges, C. J. C. *The MNIST Database of Handwritten Digits* <http://yann.lecun.com/exdb/mnist/> (1998).
48. Masquelier, T., Guyonnet, R. & Thorpe, S. J. Competitive STDP-based spike pattern learning. *Neural Comput.* **21**, 1259–1276 (2009).
49. Diehl, P. U. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* **9**, 99 (2015).
- This is a good introduction to implementing spiking neural networks with unsupervised STDP-based learning for real-world tasks such as digit recognition.**
50. Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J. & Masquelier, T. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Netw.* **99**, 56–67 (2018).
51. Neftci, E., Das, S., Pedroni, B., Kreutz-Delgado, K. & Cauwenberghs, G. Event-driven contrastive divergence for spiking neuromorphic systems. *Front. Neurosci.* **7**, 272 (2014).
52. Stromatias, E., Soto, M., Serrano-Gotarredona, T. & Linares-Barranco, B. An event-driven classifier for spiking neural networks fed with synthetic or dynamic vision sensor data. *Front. Neurosci.* **11**, 350 (2017).
53. Lee, C., Panda, P., Srinivasan, G. & Roy, K. Training deep spiking convolutional neural networks with STDP-based unsupervised pre-training followed by supervised fine-tuning. *Front. Neurosci.* **12**, 435 (2018).
54. Mostafa, H., Ramesh, V. & Cauwenberghs, G. Deep supervised learning using local errors. *Front. Neurosci.* **12**, 608 (2018).
55. Neftci, E. O., Augustine, C., Paul, S. & Deterakis, G. Event-driven random back-propagation: enabling neuromorphic deep learning machines. *Front. Neurosci.* **11**, 324 (2017).
56. Srinivasan, G., Sengupta, A. & Roy, K. Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning. *Sci. Rep.* **6**, 29545 (2016).
57. Tavanaei, A., Masquelier, T. & Maida, A. S. Acquisition of visual features through probabilistic spike-timing-dependent plasticity. In *Int. Joint Conf. on Neural Networks* 307–314 (IEEE, 2016).
58. Bagheri, A., Simeone, O. & Rajendran, B. Training probabilistic spiking neural networks with first-to-spike decoding. In *Int. Conf. on Acoustics, Speech and Signal Processing* 2986–2990 (IEEE, 2018).
59. Rastegari, M., Ordonez, V., Redmon, J. & Farhadi, A. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *Eur. Conf. on Computer Vision* 525–542 (Springer, 2016).
60. Courbariaux, M., Bengio, Y. & David, J.-P. BinaryConnect: training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems* Vol. 28 (eds Cortes, C. et al) 3123–3131 (Neural Information Processing Systems Foundation, 2015).
61. Stromatias, E. et al. Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms. *Front. Neurosci.* **9**, 222 (2015).
62. Florian, R. V. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput.* **19**, 1468–1502 (2007).
63. Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W. & Gerstner, W. Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. *PLOS Comput. Biol.* **5**, e1000586 (2009).
64. Zuo, F. et al. Habituation-based synaptic plasticity and organismic learning in a quantum perovskite. *Nat. Commun.* **8**, 240 (2017).
65. Masquelier, T. & Thorpe, S. J. Unsupervised learning of visual features through spike-timing-dependent plasticity. *PLOS Comput. Biol.* **3**, e31 (2007).
66. Rao, R. P. & Sejnowski, T. J. Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Comput.* **13**, 2221–2237 (2001).
67. Roy, S. & Basu, A. An online unsupervised structural plasticity algorithm for spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 900–910 (2017).
68. Maass, W. Liquid state machines: motivation, theory, and applications. In *Computability in Context: Computation and Logic in the Real World* (eds Cooper, S. B. & Sorbi, A.) 275–296 (Imperial College Press, 2011).
69. Schrauwen, B., D’Haene, M., Verstraeten, D. & Van Campenhout, J. Compact hardware liquid state machines on FPGA for real-time speech recognition. *Neural Netw.* **21**, 511–523 (2008).
70. Verstraeten, D., Schrauwen, B., Strooband, D. & Van Campenhout, J. Isolated word recognition with the liquid state machine: a case study. *Inf. Process. Lett.* **95**, 521–528 (2005).
71. Panda, P. & Roy, K. Learning to generate sequences with combination of Hebbian and non-Hebbian plasticity in recurrent spiking neural networks. *Front. Neurosci.* **11**, 693 (2017).
72. Maher, M. A. C., Deweerth, S. P., Mahowald, M. A. & Mead, C. A. Implementing neural architectures using analog VLSI circuits. *IEEE Trans. Circ. Syst.* **36**, 643–652 (1989).
73. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
- This seminal work established neuromorphic electronic systems as a new paradigm in hardware computing and highlights Mead’s vision of going beyond the precise and well defined nature of digital computing towards brain-like aspects.**
74. Mead, C. A. Neural hardware for vision. *Eng. Sci.* **50**, 2–7 (1987).
75. NVIDIA Launches the World’s First Graphics Processing Unit GeForce 256. https://www.nvidia.com/object/IO_20020111_5424.html (Nvidia, 1999).
76. Nageswaran, J. M., Dutt, N., Krichmar, J. L., Nicolau, A. & Veidenbaum, A. V. A configurable simulation environment for the efficient simulation of large-scale spiking neural networks on graphics processors. *Neural Netw.* **22**, 791–800 (2009).
77. Fidjeland, A. K. & Shanahan, M. P. Accelerated simulation of spiking neural networks using GPUs. In *Int. Joint Conf. on Neural Networks* 3041–3048 (IEEE, 2010).
78. Davies, M. et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).
79. Blouw, P., Choo, X., Hunsberger, E. & Eliasmith, C. Benchmarking keyword spotting efficiency on neuromorphic hardware. In *Proc. 7th Annu. Neuro-inspired Computational Elements Workshop 1* (ACM, 2018).
80. Hsu, J. How IBM got brainlike efficiency from the TrueNorth chip. *IEEE Spectrum* <https://spectrum.ieee.org/computing/hardware/how-ibm-got-brainlike-efficiency-from-the-truenorth-chip> (29 September 2014).
81. Khan, M. M. et al. SpiNNaker: mapping neural networks onto a massively parallel chip multiprocessor. In *Int. Joint Conf. on Neural Networks* 2849–2856 (IEEE, 2008).
- This was one of the first works to implement a large-scale spiking neural network on hardware using event-driven computations and commercial processors.**
82. Benjamin, B. V. et al. Neurogrid: a mixed-analog–digital multichip system for large-scale neural simulations. *Proc. IEEE* **102**, 699–716 (2014).
83. Schemmel, J. et al. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Int. Symp. Circuits and Systems* 1947–1950 (IEEE, 2010).
84. Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
- This work describes TrueNorth, the first digital custom-designed, large-scale neuromorphic processor, an outcome of the DARPA SyNAPSE programme; it was geared towards solving commercial applications through a digital neuromorphic implementation.**
85. Furber, S. Large-scale neuromorphic computing systems. *J. Neural Eng.* **13**, 051001 (2016).
86. Qiao, N. et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Front. Neurosci.* **9**, 141 (2015).
87. Indiveri, G. et al. Neuromorphic silicon neuron circuits. *Front. Neurosci.* **5**, 73 (2011).
88. Seo, J.-s. et al. A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In *Custom Integrated Circuits Conf.* 311–334 (IEEE, 2011).
89. Boahen, K. A. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Trans. Circuits Syst. II* **47**, 416–434 (2000).
- This paper describes the fundamentals of address event representation and its application to neuromorphic systems.**
90. Serrano-Gotarredona, R. et al. AER building blocks for multi-layer multi-chip neuromorphic vision systems. In *Advances in Neural Information Processing Systems* Vol. 18 (eds Weiss, Y., Schölkopf, B. & Platt, J. C.) 1217–1224 (Neural Information Processing Systems Foundation, 2006).
91. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).

92. Waldrop, M. M. The chips are down for Moore's law. *Nature* **530**, 144 (2016).
93. von Neumann, J. First draft of a report on the EDVAC. *IEEE Ann. Hist. Comput.* **15**, 27–75 (1993).
94. Mahapatra, N. R. & Venkatrao, B. The processor–memory bottleneck: problems and solutions. *Crossroads* **5**, 2 (1999).
95. Gokhale, M., Holmes, B. & Iobst, K. Processing in memory: the Terasys massively parallel PIM array. *Computer* **28**, 23–31 (1995).
96. Elliott, D., Stumm, M., Snelgrove, W. M., Cojocar, C. & McKenzie, R. Computational RAM: implementing processors in memory. *IEEE Des. Test Comput.* **16**, 32–41 (1999).
97. Ankit, A., Sengupta, A., Panda, P. & Roy, K. RESPARC: a reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking neural networks. In *Proc. 54th ACM/EDAC/IEEE Annual Design Automation Conf.* 63.2 (IEEE, 2017).
98. Bez, R. & Pirovano, A. Non-volatile memory technologies: emerging concepts and new materials. *Mater. Sci. Semicond. Process.* **7**, 349–355 (2004).
99. Xue, C. J. et al. Emerging non-volatile memories: opportunities and challenges. In *Proc. 9th Int. Conf. on Hardware/Software Codesign and System Synthesis* 325–334 (IEEE, 2011).
100. Wong, H.-S. P. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol.* **10**, 191 (2015); correction **10**, 660 (2015).
101. Chi, P. et al. Prime: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In *Proc. 43rd Int. Symp. Computer Architecture* 27–39 (IEEE, 2016).
102. Shafiee, A. et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *Proc. 43rd Int. Symp. Computer Architecture* 14–26 (IEEE, 2016).
103. Burr, G. W. et al. Neuromorphic computing using non-volatile memory. *Adv. Phys. X* **2**, 89–124 (2017).
104. Snider, G. S. Spike-timing-dependent learning in memristive nanodevices. In *Proc. Int. Symp. on Nanoscale Architectures* 85–92 (IEEE, 2008).
105. Chua, L. Memristor—the missing circuit element. *IEEE Trans. Circuit Theory* **18**, 507–519 (1971).
- This was the first work to conceptualize memristors as fundamental passive circuit elements; they are currently being investigated as high-density storage devices through various emerging technologies for conventional general-purpose and neuromorphic computing architectures.**
106. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
107. Waser, R., Dittmann, R., Staikov, G. & Szot, K. Redox-based resistive switching memories—nanioionic mechanisms, prospects, and challenges. *Adv. Mater.* **21**, 2632–2663 (2009).
108. Burr, G. W. et al. Recent progress in phase-change memory technology. *IEEE J. Em. Sel. Top. Circuits Syst.* **6**, 146–162 (2016).
109. Hosomi, M. et al. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-RAM. In *Int. Electron Devices Meeting* 459–462 (IEEE, 2005).
110. Ambrogio, S. et al. Statistical fluctuations in HfO₂ resistive-switching memory. Part I—set/reset variability. *IEEE Trans. Electron Dev.* **61**, 2912–2919 (2014).
111. Fantini, A. et al. Intrinsic switching variability in HfO₂ RRAM. In *5th Int. Memory Workshop* 30–33 (IEEE, 2013).
112. Merrikh-Bayat, F. et al. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 4782–4790 (2017).
113. Ramakrishnan, S., Hasler, P. E. & Gordon, C. Floating-gate synapses with spike-time-dependent plasticity. *IEEE Trans. Biomed. Circuits Syst.* **5**, 244–252 (2011).
114. Hasler, J. & Marr, H. B. Finding a roadmap to achieve large neuromorphic hardware systems. *Front. Neurosci.* **7**, 118 (2013).
115. Hasler, P. E., Diorio, C., Minch, B. A. & Mead, C. Single transistor learning synapses. In *Advances in Neural Information Processing Systems* Vol. 7 (eds Tesauro, G., Touretzky, D. S. & Leen, T. K.) 817–824 (Neural Information Processing Systems Foundation, 1995).
- This was one of the first works to use a non-volatile memory device—specifically, a floating-gate transistor—as a synaptic element.**
116. Holler, M., Tam, S., Castro, H. & Benson, R. An electrically trainable artificial neural network (ETANN) with 10240 ‘floating gate’ synapses. In *Int. Joint Conf. on Neural Networks* Vol. 2, 191–196 (1989).
117. Chen, P.-Y. et al. Technology-design co-optimization of resistive cross-point array for accelerating learning algorithms on chip. In *Proc. Eur. Conf. on Design, Automation & Testing* 854–859 (IEEE, 2015).
118. Chakraborty, I., Roy, D. & Roy, K. Technology aware training in memristive neuromorphic systems for nonideal synaptic crossbars. *IEEE Trans. Em. Top. Comput. Intell.* **2**, 335–344 (2018).
119. Alibart, F., Gao, L., Hoskins, B. D. & Strukov, D. B. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. *Nanotechnology* **23**, 075201 (2012).
120. Dong, Q. et al. A 4 + 2T SRAM for searching and in-memory computing with 0.3-V V_{DDmin} . *IEEE J. Solid-State Circuits* **53**, 1006–1015 (2018).
121. Agrawal, A., Jaiswal, A., Lee, C. & Roy, K. X-SRAM: enabling in-memory Boolean computations in CMOS static random-access memories. *IEEE Trans. Circuits Syst. I* **65**, 4219–4232 (2018).
122. Eckert, C. et al. Neural cache: bit-serial in-cache acceleration of deep neural networks. In *Proc. 45th Ann. Int. Symp. Computer Architecture* 383–396 (IEEE, 2018).
123. Gonugondla, S. K., Kang, M. & Shanbhag, N. R. A variation-tolerant in-memory machine-learning classifier via on-chip training. *IEEE J. Solid-State Circuits* **53**, 3163–3173 (2018).
124. Biswas, A. & Chandrakasan, A. P. Conv-RAM: an energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications. In *Int. Solid-State Circuits Conf.* 488–490 (IEEE, 2018).
125. Kang, M., Keel, M.-S., Shanbhag, N. R., Eilert, S. & Cuiwewitz, K. An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM. In *Int. Conf. on Acoustics, Speech and Signal Processing* 8326–8330 (IEEE, 2014).
126. Seshadri, V. et al. RowClone: fast and energy-efficient in-DRAM bulk data copy and initialization. In *Proc. 46th Ann. IEEE/ACM Int. Symp. Microarchitecture* 185–197 (ACM, 2013).
127. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
128. Sebastian, A. et al. Temporal correlation detection using computational phase-change memory. *Nat. Commun.* **8**, 1115 (2017).
129. Jain, S., Ranjan, A., Roy, K. & Raghunathan, A. Computing in memory with spin-transfer torque magnetic RAM. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **26**, 470–483 (2018).
130. Jabri, M. & Flower, B. Weight perturbation: an optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayer networks. *IEEE Trans. Neural Netw.* **3**, 154–157 (1992).
131. Diorio, C., Hasler, P., Minch, B. A. & Mead, C. A. A floating-gate MOS learning array with locally computed weight updates. *IEEE Trans. Electron Dev.* **44**, 2281–2289 (1997).
132. Bayat, F. M., Prezioso, M., Chakrabarti, B., Kataeva, I. & Strukov, D. Memristor-based perceptron classifier: increasing complexity and coping with imperfect hardware. In *Proc. 36th Int. Conf. on Computer-Aided Design* 549–554 (IEEE, 2017).
133. Guo, X. et al. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology. In *Int. Electron Devices Meeting* 6.5 (IEEE, 2017).
134. Liu, C., Hu, M., Strachan, J. P. & Li, H. Rescuing memristor-based neuromorphic design with high defects. In *Proc. 54th ACM/EDAC/IEEE Design Automation Conf.* 76.6 (IEEE, 2017).
135. Tuma, T., Pantazi, A., Le Gallo, M., Sebastian, A. & Eleftheriou, E. Stochastic phase-change neurons. *Nat. Nanotechnol.* **11**, 693–699 (2016).
136. Fukushima, A. et al. Spin dice: a scalable truly random number generator based on spintronics. *Appl. Phys. Express* **7**, 083001 (2014).
137. Le Gallo, M. et al. Mixed-precision in-memory computing. *Nature Electron.* **1**, 246 (2018).
138. Krstic, M., Grass, E., Gürkaynak, F. K. & Vivet, P. Globally asynchronous, locally synchronous circuits: overview and outlook. *IEEE Des. Test Comput.* **24**, 430–441 (2007).
139. Choi, H. et al. An electrically modifiable synapse array of resistive switching memory. *Nanotechnology* **20**, 345201 (2009).
140. Serrano-Gotarredona, T., Masquelier, T., Prodromakis, T., Indiveri, G. & Linares-Barranco, B. STDP and STDP variations with memristors for spiking neuromorphic learning systems. *Front. Neurosci.* **7**, 2 (2013).
141. Kuzum, D., Jeyasingh, R. G., Lee, B. & Wong, H.-S. P. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano Lett.* **12**, 2179–2186 (2012).
142. Krzysteczko, P., Münchenberger, J., Schäfers, M., Reiss, G. & Thomas, A. The memristive magnetic tunnel junction as a nanoscopic synapse–neuron system. *Adv. Mater.* **24**, 762–766 (2012).
143. Vincent, A. F. et al. Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems. *IEEE Trans. Biomed. Circuits Syst.* **9**, 166–174 (2015).
144. Sengupta, A. & Roy, K. Encoding neural and synaptic functionalities in electron spin: a pathway to efficient neuromorphic computing. *Appl. Phys. Rev.* **4**, 041105 (2017).
145. Borghetti, J. et al. ‘Memristive’ switches enable ‘stateful’ logic operations via material implication. *Nature* **464**, 873–876 (2010).
146. Hu, M. et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. In *Proc. 53rd ACM/EDAC/IEEE Annual Design Automation Conf.* 21.1 (IEEE, 2016).
147. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–789 (2017).
148. Wright, C. D., Liu, Y., Kohary, K. I., Aziz, M. M. & Hicken, R. J. Arithmetic and biologically-inspired computing using phase-change materials. *Adv. Mater.* **23**, 3408–3413 (2011).
149. Le Gallo, M., Sebastian, A., Cherubini, G., Giefers, H. & Eleftheriou, E. Compressed sensing recovery using computational memory. In *Int. Electron Devices Meeting* 28.3.1 (IEEE, 2017).
150. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65** 386 (1958).
151. Bi, G. Q. & Poo, M. M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998).

Acknowledgements We thank A. Sengupta (Pennsylvania State University), A. Raychowdhury (Georgia Institute of Technology) and S. Gupta (Purdue University) for their input. The work was supported in part by the Center for Brain-inspired Computing Enabling Autonomous Intelligence (C-BRIC), a DARPA-sponsored JUMP center, the Semiconductor Research Corporation, the National Science Foundation, Intel Corporation, the DoD Vannevar Bush Fellowship, the ONR-MURI programme, and the US Army Research Laboratory and the UK Ministry of Defence under agreement number W911NF-16-3-0001.

Author contributions All authors contributed equally in devising the structure of the paper, designing the figures and writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.R.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

A wide star–black-hole binary system from radial-velocity measurements

<https://doi.org/10.1038/s41586-019-1766-2>

Received: 1 March 2019

Accepted: 28 August 2019

Published online: 27 November 2019

Jifeng Liu^{1,2,3*}, Haotong Zhang^{1*}, Andrew W. Howard⁴, Zhongrui Bai¹, Youjun Lu^{1,2}, Roberto Soria^{2,5}, Stephen Justham^{1,2,6}, Xiangdong Li^{7,8}, Zheng Zheng⁹, Tinggui Wang¹⁰, Krzysztof Belczynski¹¹, Jorge Casares^{12,13}, Wei Zhang¹, Hailong Yuan¹, Yiqiao Dong¹, Yajuan Lei¹, Howard Isaacson¹⁴, Song Wang¹, Yu Bai¹, Yong Shao^{7,8}, Qing Gao¹, Yilun Wang^{1,2}, Zexi Niu^{1,2}, Kaiming Cui^{1,2}, Chuanjie Zheng^{1,2}, Xiaoyong Mu², Lan Zhang¹, Wei Wang^{3,15}, Alexander Heger¹⁶, Zhaoxiang Qi^{1,17}, Shilong Liao¹⁷, Mario Lattanzi¹⁸, Wei-Min Gu¹⁹, Junfeng Wang¹⁹, Jianfeng Wu¹⁹, Lijing Shao²⁰, Rongfeng Shen²¹, Xiaofeng Wang²², Joel Bregman²³, Rosanne Di Stefano²⁴, Qingzhong Liu²⁵, Zhanwen Han²⁶, Tianmeng Zhang¹, Huijuan Wang¹, Juanjuan Ren¹, Junbo Zhang¹, Jujia Zhang²⁶, Xiaoli Wang²⁶, Antonio Cabrera-Lavers^{12,27}, Romano Corradi^{12,27}, Rafael Rebolo^{13,27}, Yongheng Zhao^{1,2}, Gang Zhao^{1,2}, Yaoquan Chu¹⁰ & Xiangqun Cui²⁸

All stellar-mass black holes have hitherto been identified by X-rays emitted from gas that is accreting onto the black hole from a companion star. These systems are all binaries with a black-hole mass that is less than 30 times that of the Sun^{1–4}. Theory predicts, however, that X-ray-emitting systems form a minority of the total population of star–black-hole binaries^{5,6}. When the black hole is not accreting gas, it can be found through radial-velocity measurements of the motion of the companion star. Here we report radial-velocity measurements taken over two years of the Galactic B-type star, LB-1. We find that the motion of the B star and an accompanying H α emission line require the presence of a dark companion with a mass of 68^{+11}_{-13} solar masses, which can only be a black hole. The long orbital period of 78.9 days shows that this is a wide binary system. Gravitational-wave experiments have detected black holes of similar mass, but the formation of such massive ones in a high-metallicity environment would be extremely challenging within current stellar evolution theories.

A radial-velocity monitoring campaign using the Large Aperture Multi-Object Spectroscopic Telescope⁷ (LAMOST) to discover and study spectroscopic binaries has been running since 2016, and has obtained 26 measurements each for about 3,000 targets brighter than 14 mag in the Kepler K2-0 field of the sky⁸. One of the B-type stars towards the Galactic Anti-Centre, hereafter LB-1, located at coordinates $(l, b) = (188.23526^\circ, +02.05089^\circ)$, where l is Galactic longitude and b is Galactic latitude, with a V-band magnitude of about 11.5 mag, exhibited periodic radial-velocity variation, along with a strong, broad H α emission line that is almost stationary. Subsequent GTC/OSIRIS⁹ and Keck/HIRES¹⁰ observations between 2017 December and 2018 April have confirmed the periodic variations and the prominent H α emission line with higher spectral resolution. The spectra reveal three types of

lines: stellar absorption lines with apparent periodic motion, a broad H α emission line moving in anti-phase with much smaller amplitude, and interstellar absorption lines that are time-independent (see Fig. 1).

The overall spectral shape of LB-1 suggests a B-type star characterized by prominent Balmer absorption lines without a significant Balmer jump. The metallicity, as measured from the Si II/Mg II lines, is about $(1.2 \pm 0.2)Z_\odot$ (where the solar metallicity $Z_\odot = 0.017$), consistent with that expected for a young B-type star in the Galactic plane. TLUSTY¹¹ model fitting to the high-resolution Keck spectra leads to effective temperature $T_{\text{eff}} = 18,100 \pm 820$ K and $\log g = 3.43 \pm 0.15$, where g is the surface gravity. (The H α and H β lines were excluded from the fit because of contamination from emission.) Such values of T_{eff} and $\log g$ fit stellar models¹² around the main-sequence turn-off points with mass

¹Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China. ²School of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing, China. ³WHU-NAOC Joint Center for Astronomy, Wuhan University, Wuhan, China. ⁴Department of Astronomy, Caltech, Pasadena, CA, USA. ⁵Sydney Institute for Astronomy, The University of Sydney, Sydney, New South Wales, Australia. ⁶The Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. ⁷School of Astronomy and Space Science, Nanjing University, Nanjing, China. ⁸Key Laboratory of Modern Astronomy and Astrophysics (Nanjing University), Ministry of Education, Nanjing, China. ⁹Department of Physics and Astronomy, University of Utah, Salt Lake City, UT, USA. ¹⁰CAS Key Laboratory for Research in Galaxies and Cosmology, Department of Astronomy, University of Science and Technology of China, Hefei, China. ¹¹Nicolaus Copernicus Astronomical Centre, Polish Academy of Sciences, Warsaw, Poland. ¹²Instituto de Astrofísica de Canarias, La Laguna, Spain. ¹³Departamento de Astrofísica, Universidad de La Laguna, Santa Cruz de Tenerife, Spain. ¹⁴Astronomy Department, University of California, Berkeley, CA, USA. ¹⁵School of Physics and Technology, Wuhan University, Wuhan, China. ¹⁶Monash Centre for Astrophysics, School of Physics and Astronomy, Monash University, Victoria, Australia. ¹⁷Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai, China. ¹⁸INAF-Osservatorio Astrofisico di Torino, Pino Torinese, Italy. ¹⁹Department of Astronomy, Xiamen University, Xiamen, China. ²⁰Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing, China. ²¹School of Physics and Astronomy, Sun Yat-Sen University, Zhuhai, China. ²²Physics Department and Tsinghua Center for Astrophysics, Tsinghua University, Beijing, China. ²³Department of Astronomy, University of Michigan, Ann Arbor, MI, USA. ²⁴Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA. ²⁵Key Laboratory of Dark Matter and Space Astronomy, Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing, China. ²⁶Key Laboratory for the Structure and Evolution of Celestial Objects, Yunnan Observatories, Chinese Academy of Sciences, Kunming, China. ²⁷GRANTECAN, Breña Baja, Spain. ²⁸Nanjing Institute of Astronomical Optics and Technology, Chinese Academy of Sciences, Nanjing, China. *e-mail: jfliu@nao.cas.cn; htzhang@bao.ac.cn

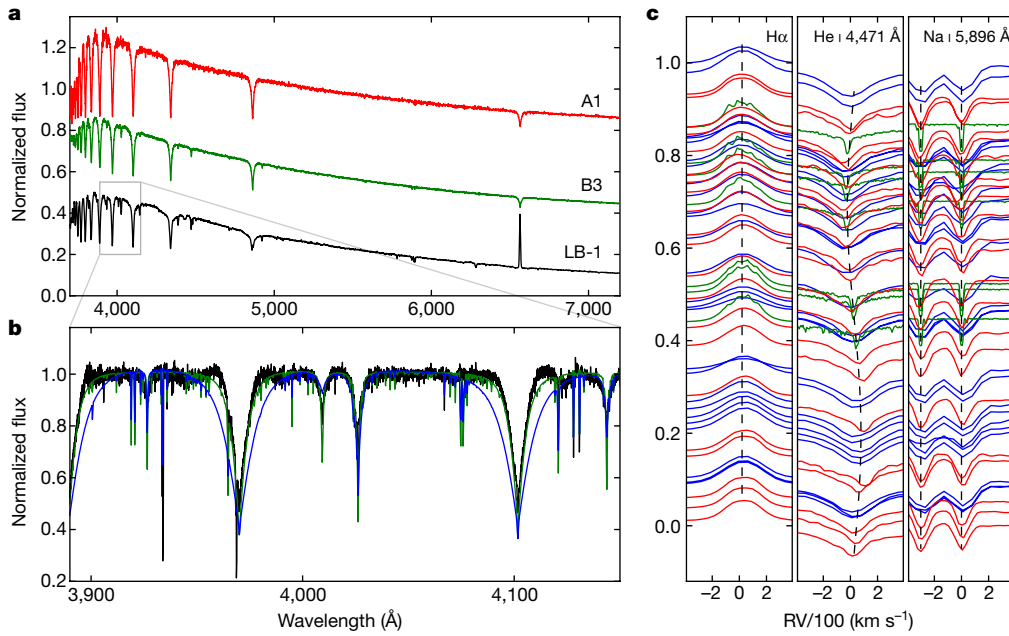


Fig. 1 | Optical spectra of LB-1. **a**, LAMOST spectrum (thin black trace; $R \approx 1,800$) with stellar templates (A1, red; B3, green; offset for clarity) overplotted. **b**, Keck/HIRES spectrum of the wavelength range boxed in **a** (black trace; $R \approx 60,000$) with the best TLUSTY model (green trace; $T_{\text{eff}} = 18,100$ K, $\log g = 3.43$, $Z = Z_{\odot}$, $v \sin i = 10$ km s $^{-1}$) overplotted. The 90% confidence level (CL) errors for the model are $\Delta T_{\text{eff}} = 820$ K and $\Delta \log g = 0.15$. Also overplotted is a comparison model with $\log g = 4.75$ (blue), which is the

highest $\log g$ of the model grid but still lower than the typical value ($\log g > 5$) for a B subdwarf. The Balmer absorption lines from this model are much wider than the observed profiles. **c**, Phased line profiles from LAMOST (blue), GTC (red) and Keck (green) observations of the H α emission line (left), the He I absorption line at $\lambda = 4,471$ Å ('He I 4,471') of the visible star (middle), and interstellar Na I absorption lines (right). The dashed lines are plotted to guide the eye. The binary phase ϕ is for the period of $P = 78.9$ d.

$M_B = 8.2^{+0.9}_{-1.2} M_{\odot}$ (where M_{\odot} is the solar mass), radius $R_B = (9 \pm 2) R_{\odot}$ (where R_{\odot} is the solar radius) and age $t = 35^{+13}_{-7}$ Myr. The best-fit model is a subgiant B-type star about 0.2 Myr after the main-sequence turn-off point. Its distance D and extinction $E(B - V)$, where B and V are respectively B-band and V-band magnitude, can be derived simultaneously from fitting its wide-band spectral energy distribution (SED), resulting in $D = 4.23 \pm 0.24$ kpc and $E(B - V) = 0.55 \pm 0.03$ mag (see Methods). These values are consistent with the 3D extinction map¹³ along LB-1's direction, so supporting this model. A subdwarf star with a similar temperature is strongly ruled out by the narrow Balmer lines, as shown in Fig. 1a, and also by the SED fitting.

The radial motion of the star, as measured from the stellar absorption lines in 26 LAMOST, 21 GTC and 7 Keck observations obtained over two years, can be best fitted with a period of $P = 78.9 \pm 0.3$ d (see Methods). Fitting a binary orbit to the folded radial-velocity curve (see Fig. 2) yields a semi-amplitude $K_B = 52.8 \pm 0.7$ km s $^{-1}$, an eccentricity $e = 0.03 \pm 0.01$, and a centre-of-mass velocity $V_{\text{OB}} = 28.7 \pm 0.5$ km s $^{-1}$. For this binary with a nearly circular orbit, the mass function is $PK_B^3/2\pi G = (1.20 \pm 0.05) M_{\odot}$ (where G is the gravitational constant), which is the absolute lower limit for the mass of the dark companion to the B star. Given that M_B is already known, the minimum mass of the dark primary can be calculated to be $6.3^{+0.4}_{-1.0} M_{\odot}$ for the edge-on geometry with $i = 90^\circ$. It must be a black hole, because a $6 M_{\odot}$ main-sequence star is only about 4–6 times fainter than the B star, and the line features would be easily detected from the Keck spectra. The black-hole mass (in M_{\odot}) will be 7.8/20/84/245 for lower inclinations at $i = 60^\circ/30^\circ/15^\circ/10^\circ$, respectively. The binary separation is about 0.9–2.3 AU for a black-hole mass of $(6\text{--}250) M_{\odot}$, making it a black-hole binary wider than any previously known Galactic black-hole binaries^{1,2}.

The prominent H α emission line is too broad, with a full-width at half-maximum of 240 km s $^{-1}$, to arise from an interloper M dwarf or surrounding nebulae, nor can it be associated with a background AGN/QSO, because this would have other prominent lines at non-zero redshift. Its complicated multi-peak profile (see Fig. 2) suggests an origin

from a gaseous Keplerian disk, which can be around the B star, the black hole, or the binary. However, the inferred gaseous disk cannot be around the B star, because the H α emission line is not tracing the motion of the B star, as clearly shown in Fig. 1b. The line profile is distinctly different from a simple double-horned profile for a Keplerian disk viewed at high inclinations. It shows a 'wine-bottle' shape with multiple peaks in the line centre, which correspond to substantial non-coherent scattering components from a disk viewed at low inclination^{14,15}. A circumbinary disk would have an inner radius truncated at 1.7 times the binary separation¹⁶, and its corresponding projected velocity is $\frac{1}{\sqrt{1.7}} \approx 0.75$ times that of the visible star, that is, about 40 km s $^{-1}$. The emission line from such a circumbinary disk will be confined to within ± 40 km s $^{-1}$, yet the observed line is three times wider with wings extended beyond ± 300 km s $^{-1}$. This supports the H α emission line not coming from a circumbinary disk, but from a disk around the black hole.

The black-hole mass can be obtained directly, using the H α emission line to trace the motion of the black hole, and comparing it to the motion of the visible star. The radial velocities of the H α line, after folding with the period of 78.9 d, can be fitted with a sinusoid in anti-phase with the visible star. However, the line centre may contain contributions from circumbinary materials and accretion spots that are not symmetrically centred on the black hole, which would decrease the inferred black-hole motion and should be masked out. We experimented with different masking schemes, and found that unmasked line wings below 1/3 height can effectively avoid contamination from the line centre, yielding a semi-amplitude $K_{\alpha} = 6.4 \pm 0.8$ km s $^{-1}$ and a centre-of-mass velocity $V_{\alpha} = 28.9 \pm 0.6$ km s $^{-1}$ (see Fig. 2a and Methods). Note that V_{α} is always consistent with V_{OB} in different schemes, confirming that the H α emission is indeed associated with the B-star–black-hole binary. The black-hole mass M_{BH} can then be estimated as $M_{\text{BH}}/M_B = K_B/K_{\alpha}$, resulting in $M_{\text{BH}} = 68^{+11}_{-13} M_{\odot}$ (with 90% errors derived from the measurement uncertainties on K_B , K_{α} and M_B). Such a black-hole mass corresponds to an inclination of $i \approx 15^\circ\text{--}18^\circ$, fully consistent with the wine-bottle shape of the H α emission line.

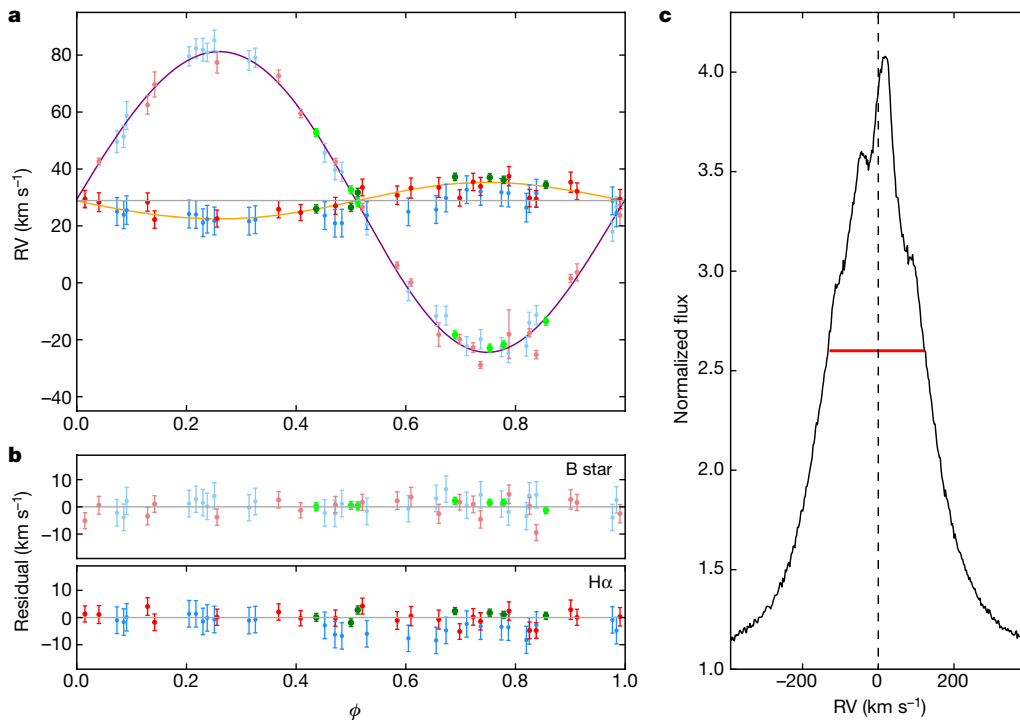


Fig. 2 | Radial motions of the visible star and the dark primary. **a**, Folded radial-velocity (RV) curves and binary orbital fits for the star and the dark primary as probed by the H α emission line. The observed data (filled circles) are from LAMOST (blue), GTC (red) and Keck (green). The error bars are the quadratic sum of the wavelength calibration uncertainty and the measurement error. The best-fit binary orbit model for the star (purple) has parameters $K_B = 52.8 \pm 0.7$ km s⁻¹, $e = 0.03 \pm 0.01$, and $V_0 = 28.7 \pm 0.5$ km s⁻¹ with a reduced χ^2 of 2.0. The best-fit model for the H α emission line (orange) has parameters

$K_\alpha = 6.4 \pm 0.8$ km s⁻¹ and $V_{0\alpha} = 28.9 \pm 0.6$ km s⁻¹ with a reduced χ^2 of 0.8. The errors quoted here are for 90% CL. The grey line with $V_0 = 28.8$ km s⁻¹ is plotted to guide the eye. **b**, Residuals for the binary orbital fits to the star (top) and to the H α emission line (bottom). The error bars are calculated as above. **c**, Representative H α emission line profile from one Keck spectrum with high spectral resolution ($R \approx 60,000$). The wine-bottle shape is caused by non-coherent scattering broadening for a disk viewed nearly pole-on. The red line represents an FWHM of about 240 km s⁻¹.

The LIGO/Virgo experiments have revealed black holes with masses of several tens of solar masses^{17,18}, much higher than previously known Galactic black holes^{1,2}. The discovery of a $70M_\odot$ black hole in LB-1 would confirm their existence in our Milky Way. However, while massive stellar black holes are expected to predominantly form in low-metallicity (that is, $<0.2Z_\odot$) environments^{19,20}, LB-1 has a B-star companion with solar metallicity. This would strongly challenge current stellar evolution models, which only allow for the formation of black holes up to $25M_\odot$ at solar metallicity^{21–23}. Formation of more-massive black holes would require reducing mass loss rates substantially at solar metallicity, and even require overcoming the well-accepted pair-instability pulsations that severely limit black-hole masses (see Methods). These strongly expected limits may suggest that the black hole in LB-1 was not formed from the collapse of only one star. One alternative is that LB-1 was initially a triple system, in which the observed B star was the outermost, least-massive component, and the present black hole was formed by the initial inner binary. Potentially, a $70M_\odot$ black hole could be formed after a ‘normal’ stellar-mass black hole merges into the core of a $\geq 60M_\odot$ star during common-envelope evolution, followed by the accretion of the massive star onto its black-hole core (see Methods). An exciting possibility is that the dark mass still contains two black holes, orbiting each other in an inner binary to which the observed star is a tertiary companion. This requires individual black-hole masses approaching $35M_\odot$, posing less of a challenge for their formation. In this case, this system would provide a laboratory to test the formation of binary black holes in triple systems.

This wide black-hole binary shows a surprisingly circular orbit that may shed light on its formation process. Circularization of such a wide binary with tidal torque would take at least a Hubble time, much longer than its age (see Methods). This rules out the possibility that LB-1 was

formed by dynamical capture of the B-type star by a black hole evolved from a low-metallicity star or by a binary black hole, as such a capture would result in an eccentric orbit that could not have been circularized by now. In the case of a co-evolving binary, this indicates a very small natal kick along with negligible mass loss when the black hole formed. Assuming an initial $e = 0$ and a symmetric mass ejection of ΔM from the black-hole progenitor, the resultant orbit will have $e = \Delta M / (M_B + M_{BH})$. Given that $e = 0.03 \pm 0.01$, ΔM must be less than 4% of the remaining mass, thus helping to form a massive black hole. Stellar evolution theories predict fallback supernova and direct black-hole formation under certain conditions, and some observations might be in favour of their existence, but direct evidence is still lacking despite observational efforts made in the past decade^{24,25}. LB-1 may be direct evidence for this process.

Our interpretation of an extraordinary $70M_\odot$ dark mass in LB-1 will be undermined if the companion mass is substantially lower than the $8M_\odot$ for the adopted B sub-giant model. To accommodate its high luminosity, we need to place the B sub-giant at a distance about twice as large as the $2.14^{+0.51}_{-0.35}$ kpc inferred from the Gaia DR2 astrometry²⁶. On one hand, this discrepancy could naturally be explained, because the binary wobble of the optical component of LB-1 is not accounted for by the Gaia DR2 single-star astrometric solution. In particular, the Gaia DR2 solution shows exceptionally large covariances, suggesting that it is unwise to simply interpret the astrometry as an accurate parallax measurement (see Methods). On the other hand, if LB-1 were indeed at that close distance, with $E(B - V) = 0.41$ mag for the appropriate line of sight at that distance¹³, its derived luminosity L would be as low as about 1/6 of the luminosity for the B sub-giant (see Methods). Taking $L \propto MT_{\text{eff}}^4/g$, and retaining the same T_{eff} and $\log g$, this implies a stellar mass M about 1/6 of our adopted value, and consequently a black-hole

mass of about $10M_{\odot}$. No natural stellar models would be consistent with such a companion, but we cannot rule out that the star is in an extreme disequilibrium state (caused, for example, by a recent outburst or supernova blast from the primary). However, the star should return to equilibrium on the Kelvin–Helmholtz timescale, which for the inferred parameters is about 10^4 yr. Thus, this low-mass companion, if it truly exists, represents a short-lived disequilibrium phase that would be extremely unlikely to be observed.

Unlike every other known stellar black hole, LB-1 has not been detected in X-ray observations. We searched for X-ray emission from this system with a 10-ks observation with the Chandra X-ray Observatory, placing an upper limit for the X-ray luminosity of $\leq 2 \times 10^{31}$ erg s $^{-1}$ (see Methods). This upper limit corresponds to about 10^{-9} of its Eddington luminosity, and suggests a mass accretion rate $\dot{M} \leq 10^{-11} M_{\odot} \text{yr}^{-1}$ for a conversion efficiency of approximately 10^{-4} at such low luminosity²⁷. Such low accretion levels could be supplied by the stellar winds of the B sub-giant²⁸. Similarly strong H α emission lines have been observed in some low-mass X-ray binaries in the X-ray quiescent state^{29,30}, where truncated accretion disks do not extend to the innermost black-hole orbits, thus preventing the emission of measurable X-ray radiation. It has long been believed that black-hole binaries in X-ray quiescence can be revealed through radial-velocity monitoring campaigns. The discovery of LB-1, with properties very unlike Galactic black-hole X-ray binaries, provides such an example. This suggests that future similar campaigns will probe a quiescent black-hole population different from the X-ray-bright one.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1766-2>.

- McClintock, J. E. & Remillard, R. A. in *Compact Stellar X-ray Sources* (eds Lewin, W. H. G. & van der Klis, M.) 157–213 (Cambridge Univ. Press, 2006).
- Casares, J. et al. A Be-type star with a black-hole companion. *Nature* **505**, 378–381 (2014).
- Silverman, J. M. & Filippenko, A. V. On IC 10 X-1, the most massive known stellar-mass black hole. *Astrophys. J.* **678**, L17–L20 (2008).
- Crowther, P. A. et al. NGC 300 X-1 is a Wolf-Rayet/black hole binary. *Mon. Not. R. Astron. Soc.* **403**, L41–L45 (2010).
- Romani, R. W. A census of low mass black hole binaries. *Astron. Astrophys.* **333**, 583–590 (1998).
- Belczynski, K. & Ziolkowski, J. On the apparent lack of Be X-ray binaries with black holes. *Astrophys. J.* **707**, 870–877 (2009).

- Cui, X.-Q. et al. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Res. Astron. Astrophys.* **12**, 1197–1242 (2012).
- Howell, S. B. et al. The K2 mission: characterization and early results. *Publ. Astron. Soc. Pacif.* **126**, 398–408 (2014).
- Cepa, J. et al. OSIRIS tunable imager and spectrograph. *Proc. SPIE* **4008**, 623–631 (2000).
- Vogt, S. S. et al. HIRES: the high-resolution echelle spectrometer on the Keck 10-m telescope. *Proc. SPIE* **2198**, 362–375 (1994).
- Hubeny, I. & Lanz, T. Non-LTE line-blanketed model atmospheres of hot stars. 1: Hybrid complete linearization/accelerated lambda iteration method. *Astrophys. J.* **439**, 875–904 (1995).
- Bressan, A. et al. PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code. *Mon. Not. R. Astron. Soc.* **427**, 127–145 (2012).
- Green, G. M. et al. A three-dimensional map of Milky Way dust. *Astrophys. J.* **810**, 25 (2015).
- Hanuschik, R. W., Hummel, W., Sutorius, E., Dietle, O. & Thimm, G. Atlas of high-resolution emission and shell lines in Be stars. Line profiles and short-term variability. *Astrophys. Space Sci.* **116**, 309–358 (1996).
- Hummel, W. Line formation in Be star envelopes I. Inhomogeneous density distributions. *Astron. Astrophys.* **289**, 458–468 (1994).
- Artymowicz, P. & Lubow, S. H. Dynamics of binary-disk interaction. 1: Resonances and disk gap sizes. *Astrophys. J.* **421**, 651–667 (1994).
- Abbott, B. P. et al. Binary black hole mergers in the first Advanced LIGO observing run. *Phys. Rev. X* **6**, 041015 (2016).
- Abbott, B. P. et al. GWTC-1: a gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs. *Phys. Rev. X* **9**, 031040 (2019).
- Belczynski, K., Holz, D. E., Bulik, T. & O’Shaughnessy, R. The first gravitational-wave source from the isolated evolution of two stars in the 40–100 solar mass range. *Nature* **534**, 512–515 (2016).
- Stevenson, S. et al. Formation of the first three gravitational-wave observations through isolated binary evolution. *Nat. Commun.* **8**, 14906 (2017).
- Belczynski, K. et al. On the maximum mass of stellar black holes. *Astrophys. J.* **714**, 1217–1226 (2010).
- Heger, A., Fryer, C. L., Woosley, S. E., Langer, N. & Hartmann, D. H. How massive single stars end their life. *Astrophys. J.* **591**, 288–300 (2003).
- Spera, M., Mapelli, M. & Bressan, A. The mass spectrum of compact remnants from the PARSEC stellar evolution tracks. *Mon. Not. R. Astron. Soc.* **451**, 4086–4103 (2015).
- Fryer, C. L. Mass limits for black hole formation. *Astrophys. J.* **522**, 413–418 (1999).
- Adams, S. M., Kochanek, C. S., Gerke, J. R., Stanek, K. Z. & Dai, X. The search for failed supernovae with the Large Binocular Telescope: confirmation of a disappearing star. *Mon. Not. R. Astron. Soc.* **468**, 4968–4981 (2017).
- Bailer-Jones, C. A. L., Rybizki, J., Fousneau, M., Mantelet, G. & Andrae, R. Estimating distance from parallaxes. IV. Distances to 1.33 billion stars in Gaia Data Release 2. *Astron. J.* **156**, 58 (2018).
- Narayan, R., Mahadevan, R. & Quataert, E. in *Theory of Black Hole Accretion Disks* (eds Abramowicz, M. A. et al.) 148–182 (Cambridge Univ. Press, 1998).
- de Jager, C., Nieuwenhuijzen, H. & van der Hucht, K. A. Mass loss rates in the Hertzsprung–Russell diagram. *Astrophys. J. Suppl. Ser.* **72**, 259–289 (1988).
- Casares, J., Charles, P. A., Naylor, T. & Pavlenko, E. P. Optical studies of V404 Cyg, the X-ray transient GS 2023 + 338 – III. The secondary star and accretion disc. *Mon. Not. R. Astron. Soc.* **265**, 834–852 (1993).
- McClintock, J. E. et al. Multiwavelength spectrum of the black hole XTE J1118 + 480 in quiescence. *Astrophys. J.* **593**, 435–451 (2003).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

Discovery and follow-up observations of LB-1

LB-1 was among the targets in the LAMOST K2-C0 time-domain survey (H.Z. et al., manuscript in preparation), which is designed to obtain time-domain spectra with the LAMOST low-resolution spectrograph ($R \approx 1,800$ over the wavelength range 3,690–9,100 Å) in a 20 square degree plate chosen from Kepler K2 Campaign 0. The plate was observed in 26 different nights from 2016 November 7 to 2018 March 23. The spectra were then reduced with the LAMOST 2D pipeline³¹.

One aim of the survey is to evaluate the binary mass function $PK^3/2\pi C = \frac{M_{\text{BH}}^3}{(M_{\text{B}} + M_{\text{BH}})^2} \sin^3 i$ once the complete radial-velocity curve can be derived from the time-domain spectral data. Here P is orbital period, K is radial-velocity semi-amplitude, and i is viewing angle. Since the mass of the brighter star in the binary can be estimated from its spectrum, the orbital eccentricity e and radial-velocity semi-amplitude K can be calculated directly from the radial-velocity curve, then the mass of the dimmer companion can be obtained immediately from the mass function given the viewing angle. About 200 out of 3,000 target stars turn out to be spectroscopic binaries with periodic radial-velocity variation. Among these, LB-1 exhibits periodic radial-velocity variations with $K = 52.8 \text{ km s}^{-1}$, $P = 78.9 \text{ d}$ and $e \approx 0$.

From the relative strength of the He I $\lambda 4,471$ versus Mg II $\lambda 4,481$ lines, we classify the star in LB-1 as a B3V star. Hot subdwarf B stars (sdBs) show spectra like B dwarfs but with much lower mass. sdBs have a shorter Balmer series ($n \approx 12$), and the He I $\lambda 4,387$ line is much weaker than the He I $\lambda 4,471$ line (ref.³²). In LB-1, the Balmer series extends to more than $n \approx 15$, which is at the blue end of the LAMOST spectral range, and the He I $\lambda 4,387$ line is clearly stronger than the He I $\lambda 4,471$ line. In addition, the LAMOST spectra show negligible N II $\lambda 3,995$ and very weak Si III $\lambda 4,552$ lines, which means LB-1 can not be a supergiant (for example, a low-mass post-AGB star). The information from the LAMOST observations is listed in Extended Data Table 1.

We carried out follow-up optical spectroscopic observations of LB-1 with GTC/OSIRIS from 2017 December 2 to 2018 April 26, using the 0.4" slit with three gratings—R2500V, R2500R and R2500I. The spectral coverage for the OSIRIS data is 450–1,000 nm, with a resolution of $\sim 3,750$. The spectra were reduced in a standard way with IRAF. After the bias subtraction and flat correction, the dispersion correction was carried out based on the line lists given in the OSIRIS manual (<http://www.gtc.iac.es/instruments/osiris/>). Raw spectra were then extracted with an aperture size of $\sim 6''$, and a standard star taken at each night was used to make the flux calibration. The wavelength calibration uncertainty is about 0.02 Å ($\sim 1.2 \text{ km s}^{-1}$). Information from the observations is listed in Extended Data Table 1.

In the period from 2017 December 9 to 2018 January 6, we observed LB-1 on seven individual nights using the Keck I telescope and the HIRES spectrometer. Exposure times range from 300 s to 600 s, and the signal-to-noise ratio (S/N) per pixel near 550 nm ranges from 80 to 120. Observations were collected using the standard California Planet Search (CPS) setup³³, resulting in a spectral resolution of $\sim 60,000$. The C2 decker ($0.87'' \times 14.0''$) was used to allow for removal of night sky line emission features and scattered moonlight. We list the information from the observations in Extended Data Table 1.

Stellar properties from high-resolution spectra

To derive the effective temperature (T_{eff}) and surface gravity ($\log g$) of LB-1, we used the spectral libraries BSTAR2006³⁴, which are based on the computer program TLUSTY¹¹. The full set of BSTAR2006 models cover T_{eff} from 15,000 K to 30,000 K with a step of 1,000 K, and $\log g$ from 1.75 to 4.75 with a step of 0.25 dex. These models include six initial metallicities; a micro-turbulence velocity of 2 km s^{-1} is adopted.

The Keck spectra are used to estimate the stellar atmosphere parameters. Using the lines Si II $\lambda 3,856$ and Si II $\lambda 5,041$, we measured the line width broadened by the stellar rotation. The program iacob_broad was

used in this step, which is available from the home-page of the IACOB project (<http://research.iac.es/proyecto/iacob/>). The $v \sin i$ is estimated as $\sim 10 \text{ km s}^{-1}$, twice that of the spectral resolution ($R \approx 60,000$) of Keck.

We performed a rotational and instrumental convolution of the original theoretical libraries to $v \sin i = 10 \text{ km s}^{-1}$ and full-width at half-maximum FWHM = 0.1 Å . Both the theoretical and observed spectra from the Keck telescope were normalized to a continuum level of unity. We used a Bayesian approach to estimate the stellar atmosphere parameters, in which each theoretical parameters is weighed by $e^{-\chi^2/2}$, where χ^2 is the goodness of fit of the model. We obtained a metallicity of $(1.18 \pm 0.18) Z_{\odot}$ with Si II $\lambda 4,131$ and $(1.17 \pm 0.11) Z_{\odot}$ with Mg II $\lambda 4,481$. Finally, using the theoretical grids with solar abundances and the observed hydrogen lines in the range of 3,750–4,150 Å, we obtained T_{eff} as $18,104 \pm 825 \text{ K}$ and $\log g$ as $3.43 \pm 0.15 \text{ dex}$. The errors were estimated using the standard deviations of the fitting results from the seven Keck spectra. These parameters prove that the optical counterpart is a B star.

Assuming solar metallicity ($Z = 0.017$), we determined the mass, radius and age of the B star. The evolutionary grid of T_{eff} and $\log g$ for stars with different initial masses were constructed on the basis of the PARSEC isochrones^{12,35} (downloaded from http://stev.oapd.inaf.it/cgi-bin/cmd_3.1). In Extended Data Fig. 1, stars located in the ellipse are considered as acceptable for the B star. We downloaded the sequences of isochrones at small steps of $\Delta(\log t) = 0.0025$, and collected the points inside the ellipse as acceptable models. Finally, we find that at $Z = 0.017$, the physical solutions (with 90% uncertainty) consistent with our constraints are: $M_{\text{B}} = 8.2^{+0.9}_{-1.2} M_{\odot}$; $R_{\text{B}} = (9 \pm 2) R_{\odot}$; $t_{\text{age}} = 35^{+13}_{-7} \text{ Myr}$.

Distance and interstellar extinction

The SED of LB-1 was extracted from the UCAC4 catalogue, 2MASS and the AllWISE data release. We used the acceptable PARSEC models to construct a grid of SEDs. By comparing the observed SEDs with the PARSEC ones, we fitted the distance and $E(B - V)$ simultaneously. Considering that the accretion disk and circumbinary materials can result in radiation in the near- and mid-infrared bands³⁶, only the U , B and V magnitudes were used in the fitting. We present the fitting results in Extended Data Fig. 2. The excesses can be found from K_s to $W4$ bands. The best fit yields the reddening value $E(B - V) = 0.55 \pm 0.03 \text{ mag}$ and the distance $4.23 \pm 0.24 \text{ kpc}$ (with 90% uncertainty). For such a distance, the Pan-STARRS 3D dust map returns an extinction of $E(B - V) \approx 0.6$, consistent with our fitting result (Extended Data Fig. 3).

The distance derived above is larger than the $2.14^{+0.51}_{-0.35} \text{ kpc}$ value from the Gaia data release 2 astrometry²⁴. This is possibly because the Gaia DR2 solution has assumed a single star for LB-1, and has mistaken the binary motion itself as part of the parallax, making the parallax and distance unreliable. In Gaia DR2, the covariances between position parameters (ra , dec) and parallax of LB-1 are much higher than other sources. The covariance dec_parallax_corr from the Gaia DR2 is -0.62 , higher than the absolute value of dec_parallax_corr of 98% of sources between 10 mag and 13 mag (G band). The covariance of ra_parallax_corr is 0.54, which is also higher than that of 96% of sources in the magnitude range.

Given the mode of operation of the Gaia astrometric instrument and the actual along scan single-CCD measurement error of 0.3 mas (ref.³⁷), an astrometric error as small as 0.1 mas per visit can be anticipated for the 11.5 magnitude of LB-1 (in each visit a star is measured on up to 9 different astrometric CCDs). Given its location, a total of about 80 of such visits are predicted throughout the Gaia 5-yr nominal mission lifetime (ending in September 2019). With these numbers in mind, there is actual hope that the 0.4 mas astrometric orbital motion can be uncovered once the single visit data are properly reduced and/or made available in the future.

The hot sdB scenario can also be rejected from the distance. The Gaia DR2 catalogue shows that³⁸ hot subluminous stars have an absolute

magnitude around 5 mag. With the G-band magnitude of 11.918 mag for LB-1, the distance estimation for an sdB would be less than 240 pc. This is seriously inconsistent with the Gaia DR2 distance and our fitting result, and is also inconsistent with the clear diffuse interstellar bands (DiBs)³⁹ in the spectra, which should be much shallower for an sdB at this distance.

As a test, we calculated the radius and mass of the B star using the observational parameters, including the V-band magnitude (-11.51 mag), the reddening value, the distance, the effective temperature, and the surface gravity. With the bolometric correction⁴⁰ being about -1.6, the bolometric magnitude of the B star is $M_{B,\text{bol}} = -4.93$ mag, and the bolometric luminosity is calculated as $L_{B,\text{bol}} = L_{\odot,\text{bol}} \times 10^{0.4(M_{\odot,\text{bol}} - M_{B,\text{bol}})} \approx 7,000 L_{\odot,\text{bol}}$. The solar bolometric magnitude and luminosity are 4.74 mag and 3.828×10^{33} erg s⁻¹, respectively. The radius is calculated as $R_B = \sqrt{\frac{L_{B,\text{bol}}}{4\pi\sigma T^4}} \approx 8.7 R_{\odot}$, and the mass is calculated as $M_B = \frac{g R_B}{G} \approx 7.5 M_{\odot}$. Both of them are consistent with the PARSEC model fitting results. The Kelvin–Helmholtz timescale $t_{\text{KH}} = \frac{GM^2}{RL}$ is defined as the time required to radiate current gravitational binding energy at its current luminosity, and represents the timescale for a star in disequilibrium to adjust back to equilibrium. In our case, t_{KH} is around 2.7×10^4 yr.

However, if we use the Gaia distance (-2.14 kpc), which corresponds to an extinction of $E(B - V) = 0.41$ from the Pan-STARRS 3D dust map, the bolometric luminosity can be estimated as $L_{B,\text{bol}} \approx 1,300 L_{\odot,\text{bol}}$. The radius and the mass would be $R_B \approx 3.6 R_{\odot}$ and $M_B \approx 1.3 M_{\odot}$, respectively. Using these parameters, the Kelvin–Helmholtz timescale would be $t_{\text{KH}} \approx 1.1 \times 10^4$ yr.

Furthermore, if we use the Gaia distance (-2.14 kpc) and assume an extinction of $E(B - V) = 0.55$ as derived from fitting the B subgiant model to the SED, the bolometric luminosity can be estimated as $L_{B,\text{bol}} \approx 1,900 L_{\odot,\text{bol}}$. The radius and the mass would be $R_B \approx 4.4 R_{\odot}$ and $M_B \approx 1.9 M_{\odot}$, respectively. The Kelvin–Helmholtz timescale is then estimated as $t_{\text{KH}} \approx 1.4 \times 10^4$ yr.

We conclude that if we place the companion at the Gaia DR2 distance, with $E(B - V)$ ranging from 0.41 mag to 0.55 mag, we will get a star in disequilibrium, with the Kelvin–Helmholtz timescale of 11,000–14,000 yr.

Radial-velocity measurements

For the B star, we measured the radial velocity by matching the model templates using the cross-correlation method. For LAMOST data, we removed the Balmer lines and DiBs, and fitted the spectrum ranging from 4,000 to 5,200 Å. For GTC and Keck data, we removed DiBs and used the spectrum ranging from 4,000 to 6,000 Å.

First, we Doppler-shifted the best theoretical spectra to a set of radial velocities. Second, we calculated the χ^2 by comparing these model spectra with the observed ones, and used the radial velocity with minimum χ^2 as the best estimation. Also, we calculated the systematic shifts between these exposures by comparing the absorption band of water vapour in the range 6,850–6,940 Å. We first used the first exposure as the reference spectra, and calculated the radial-velocity shift by cross-correlating it with the other exposures. Then we used the exposure with the median value of shift as the new reference spectra, performed the calculation again, and obtained the final systematic shifts.

One key step before radial-velocity measurements is to determine whether the H α emission is from around the black hole or from a circumbinary disk. For a circumbinary disk, the Keplerian velocity is $\sqrt{G(M_{\text{BH}} + M_B)/1.7a}$, where a is the binary separation and $1.7a$ is the typical inner radius¹⁶. The velocity of the visible star is $\sqrt{G(M_{\text{BH}} + M_B)/a_B}$, where a_B is the distance from the visible star to the barycentre ($a_B = \frac{M_{\text{BH}}}{M_B + M_{\text{BH}}} a$). The projected velocity at the inner radius of the circumbinary disk would be $\sim \frac{1}{\sqrt{1.7}} \approx 0.75$ times that of the visible star (52.8 km s^{-1}), that is, about 40 km s^{-1} . However, the observed line is three times wider with an FWHM of 240 km s^{-1} . This means that the H α emission line comes from a disk around the black hole rather than a circumbinary disk.

While it is clear that the H α emission line is associated with the black hole, it is tricky to track the black-hole motion through the H α line, because the complex structures in the line centre may be contaminated by components such as circumbinary materials, gravitationally focused accretion streams, and hot spots in the accretion disk. These components are not symmetrically centred on the black hole, hence their motion will not be in exact phase with that of the black-hole disk, and will act to decrease the black-hole motion if we include them in the calculation. Note that the line profiles can not be fitted with simple analytic forms such as Gaussian or Lorentzian profiles, so we decided instead to infer radial velocities using the barycentre of the line.

First we calculate the barycentre of the whole H α profile. The derived radial velocities over two years can be folded with the orbital period, resulting in a sinusoid with an amplitude of $1.7 \pm 0.9 \text{ km s}^{-1}$ in anti-phase with the B-star velocity. Such a line velocity, if it should represent the black-hole motion, would suggest a mass ratio of 20–67 given $M_{\text{BH}}/M_B = K_B/K_{\alpha}$, hence a black-hole mass of $(140\text{--}600)M_{\odot}$. Second we mask out the core of the line profile and calculate the barycentre from the unmasked line wings. We start by measuring the barycentre from velocity bands constrained between 1/2 FWHM up to 500 km s^{-1} , on each side of the line profile. This results in $K_{\alpha} = 4.4 \pm 0.7 \text{ km s}^{-1}$, as shown in Extended Data Table 2. This demonstrates that the line centre is indeed contaminated by components not centred on the black hole that will act to decrease the measured black-hole motion.

To explore the systematics of the derived black-hole motion, we experiment with different mask limits with inner edges in the range corresponding to 2/3 to 1/5 heights of the H α emission line, and inner edges at 120/140/170/200 km s^{-1} from the barycentre. The resulting amplitudes vary between $3.9 \pm 0.8 \text{ km s}^{-1}$ and $6.7 \pm 1.0 \text{ km s}^{-1}$, as summarized in Extended Data Table 2. It is clear from the table that the anti-phased H α motion has larger amplitude as we move away from the central part of the line, but begins to saturate after 1/3 height. This suggests that the unmasked line wings outside 1/3 height can largely avoid contamination from the line centre, and we decide to use the 1/3 height masking scheme to represent the black-hole motion, that is, $K_{\alpha} = 6.4 \pm 0.8 \text{ km s}^{-1}$ as shown in Extended Data Tables 2 and 3.

If we had many more high-resolution Keck/HIRES spectra covering all binary phases, we would be able to reconstruct the morphology of the accretion disk and other components, giving a detailed description of its asymmetric shape and size. This of course will give us a more accurate determination of the black-hole mass, and we will pursue such a (costly) follow-up campaign in the coming years. Our current LAMOST/GTC/Keck observations, however, are already enough for a rough estimate of the black-hole mass.

For LAMOST and GTC observations, there are multiple exposures during a single night. We used the averaged value as the radial velocity at that day. The measurement error was estimated using the standard deviation of multi-exposures during one night. For Keck observations, we use the measurement error, which is about 1 km s^{-1} . The system difference between days is calibrated by both the telluric emission (for LAMOST) or absorption (for GTC and Keck) lines, and also the diffuse interstellar absorption lines/bands at the Na I D lines, 5,782 Å and 6,284 Å. The uncertainty for the radial velocity is the quadratic sum of the wavelength calibration uncertainty and the measurement error.

Period and orbital parameters

Using the Lomb–Scargle^{41,42} method, we measured the period of LB-1 with the radial-velocity curve from LAMOST, GTC and Keck observations. The period is $78.9 \pm 0.3 \text{ d}$ (Extended Data Fig. 4). We fitted the radial-velocity data of the B star (54 points) and the H α line wing (54 points) simultaneously, using the equation

$$V = K[\cos(\theta + \omega) + e \cos(\omega)] + V_0 \quad (1)$$

Article

where K is the semi-amplitude of the radial-velocity curve, θ is the phase angle, ω is the longitude of periastron, and V_0 is the system velocity. The best fit parameters (that is, eccentricity e , semi-amplitude K_B and K_α , velocity V_{0B} and $V_{0\alpha}$) are listed in Extended Data Table 3. The best-fit for the B-star motion has a reduced χ^2 of 2.0, while the best-fit for the H α motion (for the 1/3 height scheme) has a reduced χ^2 of 0.8. To obtain the uncertainty of one parameter, we fixed the other parameters at the best-fit values and re-did the fitting. Then, the uncertainty of that parameter was estimated with $\Delta\chi^2 = 2.706$ (at 90% confidence) and $\Delta\chi^2 = 6.635$ (at 99% confidence).

We compared the fittings of the H α velocity using one sinusoid and one horizontal line. For the sinusoid fitting, there are two free parameters (that is, K_α and $V_{0\alpha}$); for the line fitting, there is one free parameter (that is, $V_{0\alpha}$). Therefore, the degrees of freedom for the two fittings are 52 and 53, respectively. The χ^2 for the two fittings are 109.49 and 219.24, respectively. Using the F-test, we find the sinusoid fitting is statistically significantly better than the line fitting ($P < 0.01\%$).

The separation a can be calculated from Kepler's third law $a = \left[\frac{G(M_B + M_{BH})P^2}{4\pi^2} \right]^{1/3}$ for each pair of M_{BH} and M_B . The ranges of the separation a and the semi-major axis a_B are shown in Extended Data Figs. 5 and 6 respectively under the limitations of M_B and M_{BH} , which clearly show LB-1 is a wide binary.

Black-hole formation

Individual stellar progenitor scenario. First, let us assume that the dark object in LB-1 is a single black hole formed from an individual star. Its mass depends on three major factors: (i) initial stellar mass; (ii) wind mass loss during the star's life; (iii) black-hole formation process during the final core-collapse/supernova. The initial stellar mass sets an upper limit to the black-hole mass, while winds and collapse/explosion processes are responsible for removing stellar mass and reducing the black-hole mass. All these aspects of stellar evolution are highly uncertain, which allows for a wide range of possibilities when it comes to black-hole mass calculations.

Guided by observations (or the lack thereof), we constructed a set of models based on stellar evolution calculations⁴³ to estimate the maximum black-hole mass at solar metallicity ($Z = 0.017$). We allow stars to form with initial masses as high as $200M_\odot$; at least one such star has already been discovered⁴⁴. Recent observations indicate that stellar winds may be overestimated by as much as a factor of 10 for some massive stars⁴⁵, compared with standard values⁴⁶; hence, in our calculation we reduce the theoretically predicted wind mass-loss rates by a factor of 2 to 3. At the end of a massive star's life, we allow for direct black-hole formation with no supernova explosion or associated mass loss²⁵. Such a mode of black-hole formation is supported by the low peculiar velocities observed in the most massive Galactic black holes known to date⁴⁷, and by the claimed observation of a luminous star disappearing without a supernova²⁶. Finally, we also eliminate mass loss from pair-instability pulsations during the supernova explosion; we note that, despite the large amount of theoretical work on pair-instability mass loss and supernovae, so far there is no observational evidence to support this mechanism.

We have incorporated all these options into the population synthesis code StarTrack^{48,49} to estimate the maximum black-hole mass in the Galaxy. In Extended Data Fig. 7, we present our models that challenge the currently accepted paradigms, but that can possibly explain the black-hole mass in LB-1. Our first model (magenta line) shows the standard prediction of black-hole mass as a function of initial stellar mass. Black holes form only with relatively low masses of $\sim(5-15)M_\odot$, as a result of strong stellar winds that remove most of the stellar mass before core-collapse. In our second model (blue line), we reduce stellar winds by a factor of 2. This reduction factor is applied to all types of winds: from O stars, B supergiants, luminous blue variables, and Wolf-Rayet stars²¹. As a result, the black-hole mass can reach $\sim 30M_\odot$. In our third model (red line), we reduce winds by a factor of 3. The maximum

black-hole mass is now $\sim 60M_\odot$, from a star with an initial mass of $120M_\odot$, which loses half of its mass in stellar winds before direct collapse. Stars more massive than $120M_\odot$ grow massive helium cores ($M_{He} > 45M_\odot$) and are thus subject to pair-instability pulsation supernova mass losses. Precise estimates of this type of mass loss are model-dependent, but most models agree^{50,51} that the black-hole remnants are less massive than $\sim 50M_\odot$. In our model, black holes formed after pair-instability pulsations are assumed to be always less massive than $\sim 40M_\odot$ (ref. ⁵²). In our fourth model (black line), we not only reduce stellar winds by a factor of 3, but also turn off pair-instability pulsation supernova mass losses. The maximum black-hole mass reaches $\sim 80M_\odot$, for a maximum initial stellar mass of $200M_\odot$: enough to explain the dark mass in LB-1 as a single stellar black hole.

Binary progenitor scenario. Second, let us suppose that the progenitor of LB-1 consisted of a massive (but not extraordinary) binary, with two stars of initial mass $\geq 60M_\odot$ each, and a much less massive third star (the B3 star we see today) orbiting around the O-star pair. The more massive of the two O stars evolves first, forming a black hole with a mass of $\sim(10-20)M_\odot$ at solar metallicity. If the other O star has a mass ≥ 3.5 times the black-hole mass, the system is thought to evolve through a common envelope phase^{19,53,54}. Let us then assume that the black hole sinks towards the core of the O star before the common envelope is ejected. What happens at this stage is an open question; one scenario is that the core is tidally disrupted and accreted by the in-spiralling black hole, in a regime of radiatively inefficient (advective), hyper-critical accretion⁵⁵⁻⁵⁷. If the radiative and mechanical feedback from the accreting black hole is not sufficient to destroy the star, or is collimated along the polar direction, most of the O-star envelope may end up also being accreted into the black-hole core⁵⁸, in a kind of triggered direct collapse. The final result may be a single black hole with a mass $> 60M_\odot$ with the B3 star still orbiting around it.

An alternative scenario is that the O-star plus black-hole binary system is too wide to undergo a common envelope phase, or the mass ratio is not high enough, and the system evolves instead in a slower spiral-in⁵⁴. At the end of this phase, the O star will also collapse into a black hole, without a merger. The dark mass measured in LB-1 may be a binary black hole, with masses of $\sim 35M_\odot$ for each component. The advantage of this scenario is that the formation of two $35M_\odot$ black holes from two massive stars is less problematic than the formation of a $70M_\odot$ black hole from a single star. In this scenario, too, the B3 optical counterpart is the small third component of the triple stellar system.

Circularization timescale

Tidal interactions in a binary tend to circularize its initially eccentric orbit. To estimate the circularization timescale of a B-star-black-hole binary, we use the MESA code⁵⁹ to simulate the orbital evolution. The B-star mass and the orbital period are initially set to be $8M_\odot$ and 79 d, respectively. During the evolution, we follow the evolution of the B star from the zero-age main-sequence phase to the age of 50 Myr when the star slightly evolves off the main-sequence stage. Since such a B star has a radiative envelope, we adopt the mechanism involving dynamical tides with radiative damping⁶⁰ to deal with the binary orbital evolution.

We vary the initial orbital eccentricity in the range of 0.1–0.5 and the initial black-hole mass in the range $(40-1,000)M_\odot$ to test their influence on orbital circularization. Our calculations show that the orbital eccentricity of the binary is nearly unchanged and the corresponding circularization timescale is always larger than 10^{14} yr over the whole 50 Myr. It is argued that a significant enhancement of radiative damping is required to match the observed eccentricity-period distribution in late-type binaries⁶¹, so our calculated circularization timescale may be overestimated to some extent. Since the B star (with a radius less than its current value of $(9 \pm 2)R_\odot$) is well within its Roche lobe (with a size of $(73-71)R_\odot$ corresponding to the black-hole mass of $(10-100)M_\odot$),

tides are not expected to be important, independent of the mechanism behind tidal damping.

If the masses of both components of the binary system are decreased by a factor of 6, the black hole's companion is now a low-mass ($\sim 1.3M_{\odot}$) star with a convective envelope. We then apply the mechanism involving equilibrium tide with convective damping⁶⁰ to simulate the binary orbital evolution. We find that the circularization timescale is still larger than 10^{12} yr before the low-mass star climbs to the red giant branch (corresponding to a radius of $\sim 3R_{\odot}$).

X-ray luminosity and Eddington ratio

We obtained a 10-ks DDT observation with Chandra ACIS-S3 on 2018 January 13. We reprocessed the data with ciao version 4.10; we used the ciao task *scrflux* for flux measurements. We do not detect the source, which places a 90% upper limit to the 0.5–7 keV net count rate of $\sim 3.8 \times 10^{-4}$ counts per s.

In order to convert this limit to an unabsorbed flux limit, we used a grid of plausible values of photon index and column density. The typical power-law photon index of black-hole binaries in the quiescent state is^{62,63} $\Gamma \approx 1.5\text{--}2.1$. To constrain the column density, we used the best-fitting value of the optical reddening $E(B - V) = 0.55$ mag. Applying the standard linear relation between the hydrogen column density N_{H} and the reddening⁶⁴ $N_{\text{H}} = 5.8 \times 10^{21}/E(B - V)$, we obtain $N_{\text{H}} \approx 3.2 \times 10^{21} \text{ cm}^{-2}$. A similar result ($N_{\text{H}} \approx (3.1\text{--}3.8) \times 10^{21} \text{ cm}^{-2}$) is obtained from the best-fitting relation between $A_{\text{V}} \approx 3.1E(B - V)$ and hydrogen column density^{65,66}. The line-of-sight Galactic column density in the direction of LB-1 provides a plausible upper limit⁶⁷ $N_{\text{H}} \approx 4.7 \times 10^{21} \text{ cm}^{-2}$. The saturated relation⁶⁸ provides a lower limit $N_{\text{H}} \approx 1.0 \times 10^{21} \text{ cm}^{-2}$ for $E(B - V) = 0.55$ mag. The result of our analysis over this range of photon indices and column densities is that LB-1 is not detected down to a 90% upper limit of $f_{0.3-8} < 3.9 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$ for the absorbed flux in the 0.3–8 keV band (assuming the softest slope), or $f_{0.3-8} < 4.8 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$ (assuming the hardest slope). At the adopted distance of 4.23 kpc, the 90% upper limits for the emitted luminosity are $L_{0.3-8} < 1.2 \times 10^{31} \text{ erg s}^{-1}$ (assuming the lowest limit of N_{H}), or $L_{0.3-8} < 1.8 \times 10^{31} \text{ erg s}^{-1}$ (assuming the highest value of N_{H}). Finally, for our inferred black-hole mass of $\sim 70M_{\odot}$, this corresponds to an Eddington ratio $L_{\text{X}}/L_{\text{Edd}} \lesssim 2 \times 10^{-9}$. This is the lowest value recorded for a quiescent Galactic black-hole binary^{62,69,70}, and similar to or lower than any quiescent nuclear black hole in nearby galaxies^{71–73}.

At very low accretion rates, the radiative efficiency η is reduced: a standard scaling for the ADAF model^{74,75} is $\eta \approx 10\dot{m}$ where $\dot{m} \equiv \dot{M}/\dot{M}_{\text{Edd}}$ and $\dot{M}_{\text{Edd}} \equiv L_{\text{Edd}}/(0.1c^2)$. Here \dot{M} is the accretion rate, \dot{M}_{Edd} is the Eddington accretion rate, and \dot{M} is the Eddington ratio. A similar scaling of $\eta \approx 0.7(\alpha/0.3)(L/L_{\text{Edd}})^{1/2}$ has been derived⁷⁶. An even steeper dependence of η with accretion rate ($\eta \propto \dot{m}^{1.3}$, $L \propto \dot{m}^{2.3}$) was proposed^{77,78}. Thus, our observed Eddington ratio $L_{\text{X}}/L_{\text{Edd}} \lesssim 2 \times 10^{-9}$ suggests $\dot{M} \lesssim 10^{-5}\dot{M}_{\text{Edd}} \approx 10^{-11}M_{\odot}\text{yr}^{-1}$ (with an uncertainty of a factor of 2, between alternative scaling approximations of the radiative efficiency at low accretion rates).

Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding authors upon reasonable request.

31. Bai, Z. R. et al. Sky subtraction for LAMOST. *Res. Astron. Astrophys.* **17**, 091 (2017).
32. Sargent, W. L. & Searle, L. A quantitative description of the spectra of the brighter Feige stars. *Astrophys. J.* **152**, 443–452 (1968).
33. Howard, A. W. et al. The California Planet Survey. I. Four new giant exoplanets. *Astrophys. J.* **721**, 1467–1481 (2010).
34. Lanz, T. & Hubeny, I. A grid of NLTE line-blanketed model atmospheres of early B-type stars. *Astrophys. J. Suppl. Ser.* **169**, 83–104 (2007).
35. Marigo, P. et al. A new generation of PARSEC-COLIBRI stellar isochrones including the TP-AGB phase. *Astrophys. J.* **835**, 77 (2017).
36. Bonatto, C., Bica, E., Ortolani, S. & Barbay, B. Detection of K_{c} -excess stars in the 14 Myr open cluster NGC 4755. *Astron. Astrophys.* **453**, 121–132 (2006).

37. Lindegren, L. et al. Gaia Data Release 2. The astrometric solution. *Astron. Astrophys.* **616**, A2 (2018).
38. Geier, S., Raddi, R., Gentile Fusillo, N. P. & Marsh, T. R. The population of hot subdwarf stars studied with Gaia. II. The Gaia DR2 catalogue of hot subluminescent stars. *Astron. Astrophys.* **621**, A38 (2019).
39. Friedman, S. D. et al. Studies of diffuse interstellar bands V. Pairwise correlations of eight strong DIBs and neutral hydrogen, molecular hydrogen, and color excess. *Astrophys. J.* **727**, 33 (2011).
40. Girardi, L., Grebel, E. K., Odenkirchen, M. & Chiosi, C. Theoretical isochrones in several photometric systems. II. The Sloan Digital Sky Survey ugriz system. *Astron. Astrophys.* **422**, 205–215 (2004).
41. Lomb, N. R. Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.* **39**, 447–462 (1976).
42. Scargle, J. D. Studies in astronomical time series analysis. II – Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* **263**, 835–853 (1982).
43. Hurley, J. R., Pols, O. R. & Tout, C. A. Comprehensive analytic formulae for stellar evolution as a function of mass and metallicity. *Mon. Not. R. Astron. Soc.* **315**, 543–569 (2000).
44. Crowther, P. A. et al. The R136 star cluster hosts several stars whose individual masses greatly exceed the accepted $150 M_{\odot}$ stellar mass limit. *Mon. Not. R. Astron. Soc.* **408**, 731–751 (2010).
45. Ramachandran, V. et al. Testing massive star evolution, star formation history, and feedback at low metallicity. Spectroscopic analysis of OB stars in the SMC wing. *Astron. Astrophys.* **625**, A104 (2019).
46. Vink, J. S., de Koter, A. & Lamers, H. J. G. L. M. Mass-loss predictions for O and B stars as a function of metallicity. *Astron. Astrophys.* **369**, 574–588 (2001).
47. Mirabel, I. F. & Rodrigues, I. Formation of a black hole in the dark. *Science* **300**, 1119–1120 (2003).
48. Belczynski, K. et al. Compact object modeling with the StarTrack population synthesis code. *Astrophys. J. Suppl. Ser.* **174**, 223–260 (2008).
49. Belczynski, K. et al. The evolutionary roads leading to low effective spins, high black hole masses, and O1/O2 rates of LIGO/Virgo binary black holes. Preprint at <http://arXiv.org/abs/1706.07053> (2017).
50. Woosley, S. E. Pulsational pair-instability supernovae. *Astrophys. J.* **836**, 244 (2017).
51. Leung, S.-C., Nomoto, K. & Blinnikov, S. Pulsational pair-instability supernova I: pre-collapse evolution and pulsational mass ejection. Preprint at <http://arXiv.org/abs/1901.11136> (2019).
52. Belczynski, K. et al. The effect of pair-instability mass loss on black-hole mergers. *Astron. Astrophys.* **594**, A97 (2016).
53. Dominik, M. et al. Double compact objects. I. The significance of the common envelope on merger rates. *Astrophys. J.* **759**, 52 (2012).
54. van den Heuvel, E. P. J., Portegies Zwart, S. F. & de Mink, S. E. Forming short-period Wolf-Rayet X-ray binaries and double black holes through stable mass transfer. *Mon. Not. R. Astron. Soc.* **471**, 4256–4264 (2017).
55. Jiang, Y.-F., Stone, J. M. & Davis, S. W. A global three-dimensional radiation magnetohydrodynamic simulation of super-Eddington accretion disks. *Astrophys. J.* **796**, 106 (2014).
56. Sądowski, A., Narayan, R., McKinney, J. C. & Tchekhovskoy, A. Numerical simulations of super-critical black hole accretion flows in general relativity. *Mon. Not. R. Astron. Soc.* **439**, 503–520 (2014).
57. Abramowicz, M. A. & Fragile, P. C. Foundations of black hole accretion disk theory. *Living Rev. Relativ.* **16**, 1 (2013).
58. Abubekrov, M. K., Antokhina, E. A., Bogomazov, A. I. & Cherepashchuk, A. M. The mass of the black hole in the X-ray binary M33 X-7 and the evolutionary status of M33 X-7 and IC 10 X-1. *Astron. Rep.* **53**, 232–242 (2009).
59. Paxton, B. et al. Modules for experiments in stellar astrophysics (MESA). *Astrophys. J. Suppl. Ser.* **192**, 3 (2011).
60. Hurley, J. R., Tout, C. A. & Pols, O. R. Evolution of binary stars and the effect of tides on binary populations. *Mon. Not. R. Astron. Soc.* **329**, 897–928 (2002).
61. Claret, A. New grids of stellar models including tidal-evolution constants up to carbon burning. IV. From 0.8 to 125 M: high metallicities ($Z = 0.04\text{--}0.10$). *Astron. Astrophys.* **467**, 1389–1396 (2007).
62. Plotkin, R. M., Gallo, E. & Jonker, P. G. The X-ray spectral evolution of Galactic black hole X-ray binaries toward quiescence. *Astrophys. J.* **773**, 59 (2013).
63. Remillard, R. A. & McClintock, J. E. X-ray properties of black-hole binaries. *Annu. Rev. Astron. Astrophys.* **44**, 49–92 (2006).
64. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
65. Predehl, P. & Schmitt, J. H. M. M. X-raying the interstellar medium: ROSAT observations of dust scattering halos. *Astron. Astrophys.* **293**, 889–905 (1995).
66. Güver, T. & Özel, F. The relation between optical extinction and hydrogen column density in the Galaxy. *Mon. Not. R. Astron. Soc.* **400**, 2050–2053 (2009).
67. Kalberla, P. M. W. et al. The Leiden/Argentine/Bonn (LAB) survey of Galactic HI. Final data release of the combined LDS and IAR surveys with improved stray-radiation corrections. *Astron. Astrophys.* **440**, 775–782 (2005).
68. Liszt, H. N(H I)/E(B - V). *Astrophys. J.* **780**, 10 (2014).
69. García, M. R. et al. New evidence for black hole event horizons from Chandra. *Astrophys. J.* **553**, L47–L50 (2001).
70. McClintock, J. E., Narayan, R. & Rybicki, G. B. On the lack of thermal emission from the quiescent black hole XTE J1118 + 480: evidence for the event horizon. *Astrophys. J.* **615**, 402–415 (2004).
71. Yuan, F., Yu, Z. & Ho, L. C. Revisiting the “fundamental plane” of black hole activity at extremely low luminosities. *Astrophys. J.* **703**, 1034–1043 (2009).
72. Ho, L. C. Radiatively inefficient accretion in nearby galaxies. *Astrophys. J.* **699**, 626–637 (2009).
73. Gallo, E. et al. AMUSE-Virgo. II. Down-sizing in black hole accretion. *Astrophys. J.* **714**, 25–36 (2010).

74. Narayan, R., Mahadevan, R., Grindlay, J. E., Popham, R. G. & Gammie, C. Advection-dominated accretion model of Sagittarius A*: evidence for a black hole at the Galactic center. *Astrophys. J.* **492**, 554–568 (1998).
75. Yuan, F. & Narayan, R. Hot accretion flows around black holes. *Annu. Rev. Astron. Astrophys.* **52**, 529–588 (2014).
76. Mahadevan, R. Scaling laws for advection-dominated flows: applications to low-luminosity galactic nuclei. *Astrophys. J.* **477**, 585–601 (1997).
77. Merloni, A., Heinz, S. & di Matteo, T. A fundamental plane of black hole activity. *Mon. Not. R. Astron. Soc.* **345**, 1057–1076 (2003).
78. Russell, H. R. et al. Radiative efficiency, variability and Bondi accretion on to massive black holes: the transition from radio AGN to quasars in brightest cluster galaxies. *Mon. Not. R. Astron. Soc.* **432**, 530–553 (2013).

Acknowledgements We thank D. Wang, J. Miller, E. Cackett, R. Narayan, H. Chen, B. Zhang, C. Motch, M. Bessel, G. Da Costa, A. Bogomazov, S. Wang and many others for helpful discussions. This work was supported by the National Science Foundation of China (NSFC) under grant numbers 11988101/11425313 (J.L.), 11773015/11333004/U1838201 (X.L.), 11603010 (Y.S.), 11690024 (Y. Lei), U1531118 (W.Z.), 11603035 (S.W.), 11733009 (Q.L.) and 11325313/11633002 (X.W.). It was also supported by the National Key Research and Development Program of China (NKRDPC) under grant numbers 2019YFA0405504 and 2016YFA0400804 (J.L.), 2016YFA0400803 (X.L.) and 2016YFA0400704 (Y. Lu). J.C. acknowledges support by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) under grant AYA2017-83216-P. K.B. acknowledges support from the Polish National Science Center (NCN) grants OPUS (2015/19/B/ST9/01099) and Maestro (2018/30/A/ST9/00050). This work was only made possible with LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope), a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project was provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy

of Sciences. This work is partly based on observations made with the Gran Telescopio Canarias (GTC), installed in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias, in the island of La Palma. Part of the data was obtained at the W.M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and the National Aeronautics and Space Administration. The Observatory was made possible by the generous financial support of the W.M. Keck Foundation. The scientific results reported in this article are based in part on observations made by the Chandra X-ray Observatory (ObsID 20928). This research has made use of software provided by the Chandra X-ray Center (CXC) in the application packages CIAO.

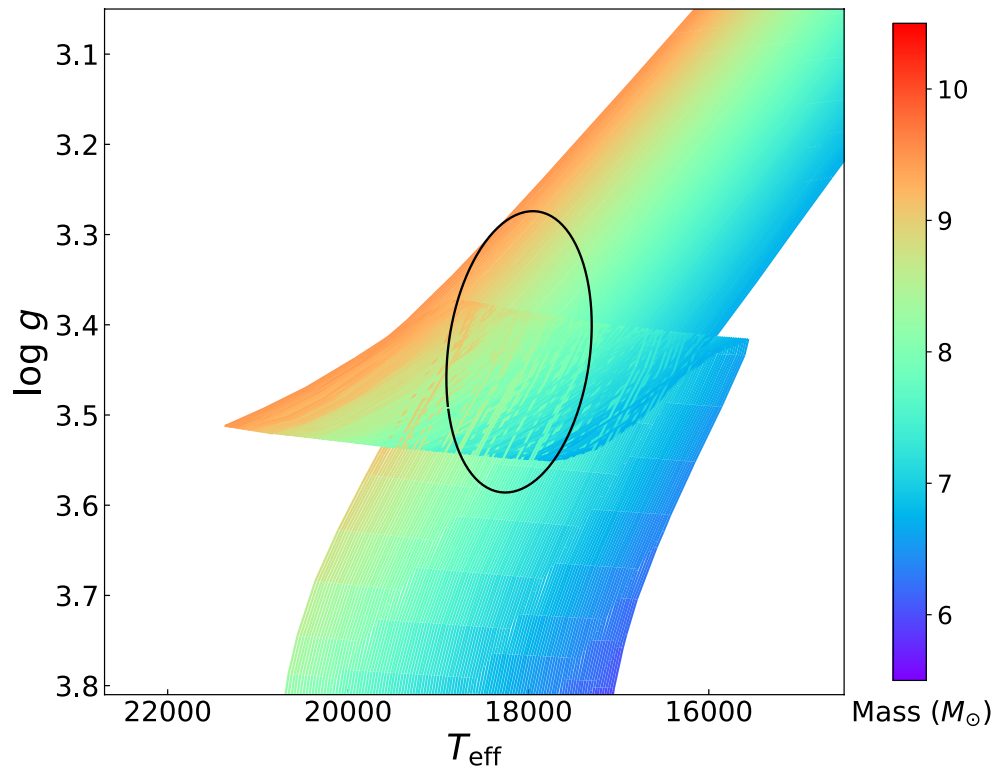
Author contributions J.L. and H.Z. are equally responsible for supervising the discovery and follow-up observations. H.Z. and Z.H. proposed the LAMOST monitoring campaign, and H.Z.'s group reduced the LAMOST data with meticulous efforts. J.L. proposed the GTC/Keck/Chandra observations, and his and H.Z.'s groups carried out subsequent data reduction and analysis. J.L. wrote the manuscript with help mainly from H.Z., Y. Lu, R.S., S.W., X.L., Y.S., T.W., Y.B., Z.B., W.Z., Q.G., Y.W., Z.Z., K.B. and J.C. W.W., A.H., W.M.G., J. Wang, J. Wu, L.S., R.S., X.W., J.B., R.D.S. and Q.L. also contributed to the physical interpretation and discussion. H.Y., Y.D., Y. Lei, Z.N., K.C., C.Z., X.M., L.Z., T.Z., H.W., J.R., Junbo Zhang, Jujia Zhang and X.W. also contributed to data collection and reduction. A.W.H. and H.I. contributed to collecting and reducing Keck data. A.C.L., R.C. and R.R. contributed to collecting and reducing GTC data. Z.Q., S.L. and M.L. contributed to utilization of Gaia data. Y.Z., G.Z., Y.C. and X.C. contributed to the implementation of LAMOST. All contributed to the paper in various forms.

Competing interests The authors declare no competing interests.

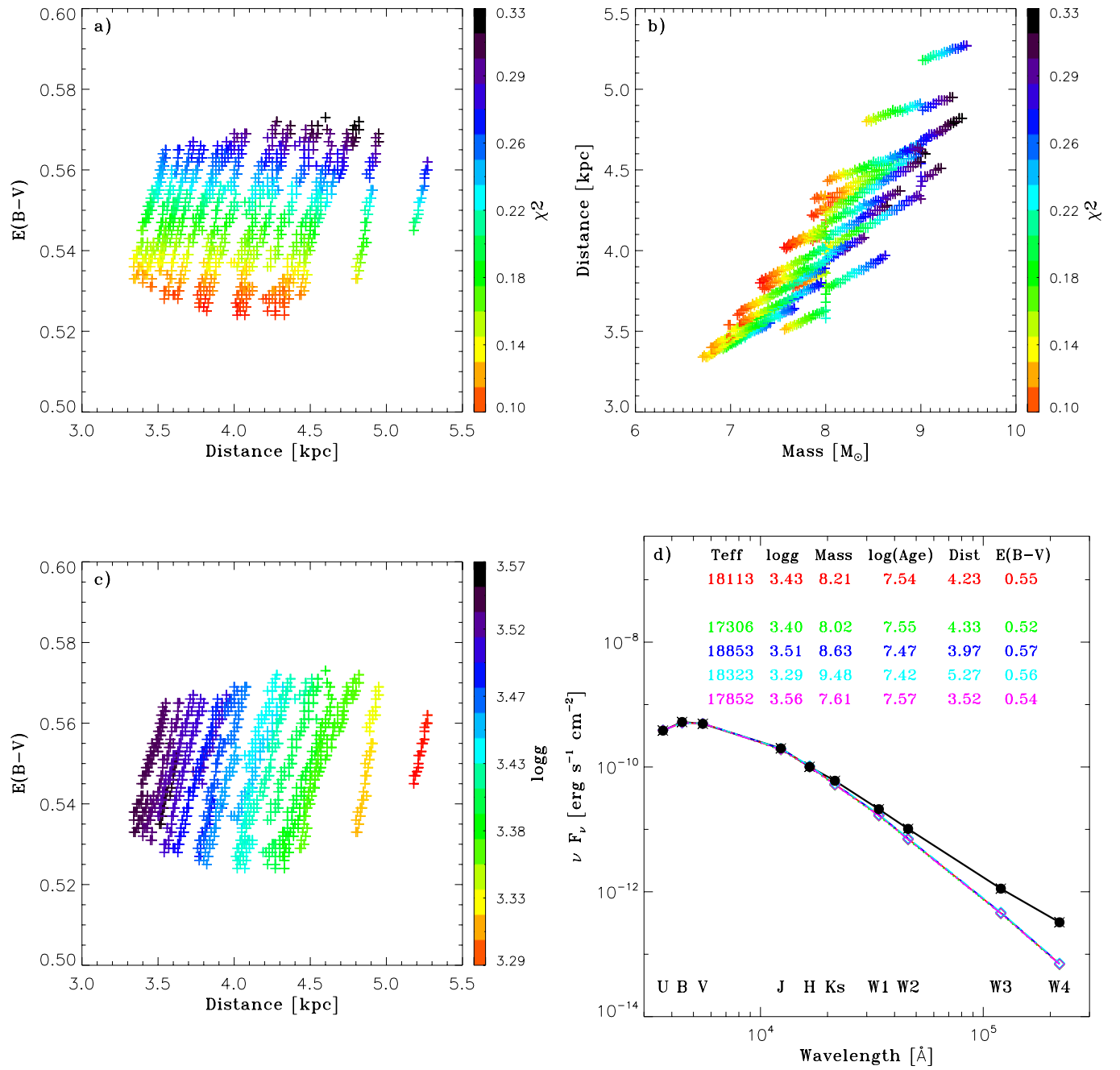
Additional information

Correspondence and requests for materials should be addressed to J.L. or H.Z.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

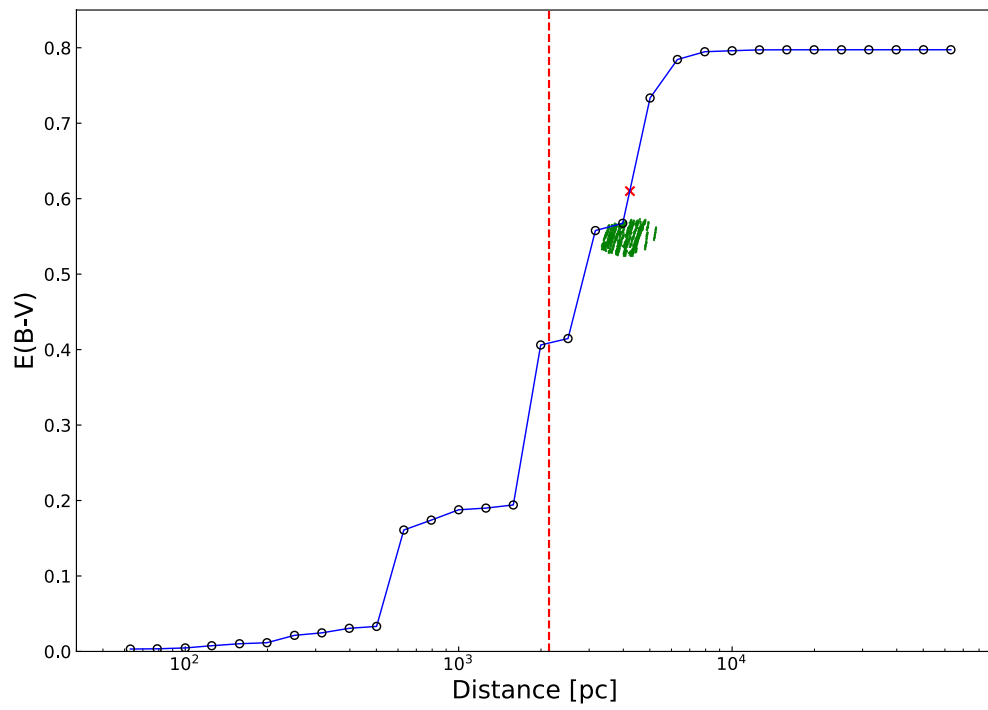


Extended Data Fig. 1 | Using isochrones from PARSEC models. The grid of $\log g$ and T_{eff} was constructed using the PARSEC isochrones. The colour bar represents initial stellar mass. The black ellipse indicates the 90% uncertainty of the T_{eff} and $\log g$ of the B star; all points inside it are considered as acceptable models for the B star.



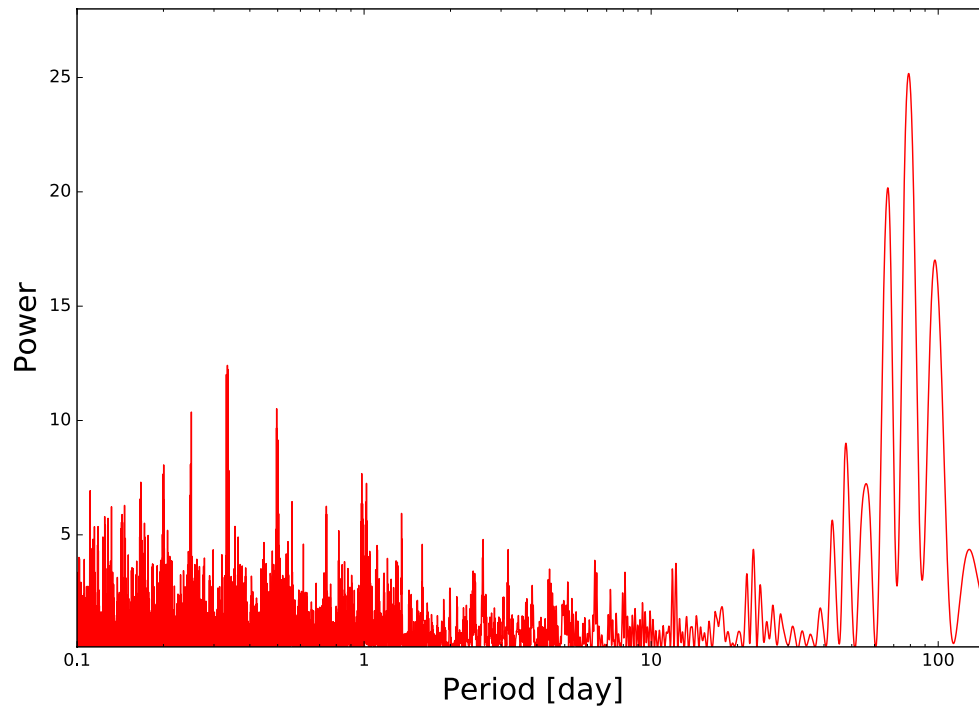
Extended Data Fig. 2 | SED fitting results for the B star. **a**, $E(B-V)$ versus distance, both of which are determined from the SED fitting. The colour bar indicates χ^2 . **b**, Distance versus stellar mass, the latter being determined from the acceptable PARSEC models of the B star. The colour bar indicates χ^2 . **c**, $E(B-V)$ versus distance. The colour bar indicates $\log g$, while the colour bar

indicates χ^2 in **a**. **d**, Several examples of the SED fitting. The black squares are the data from the UCAC4, 2MASS and AllWISE catalogues. The diamonds with different colours indicate magnitudes from different models. See Methods for details.

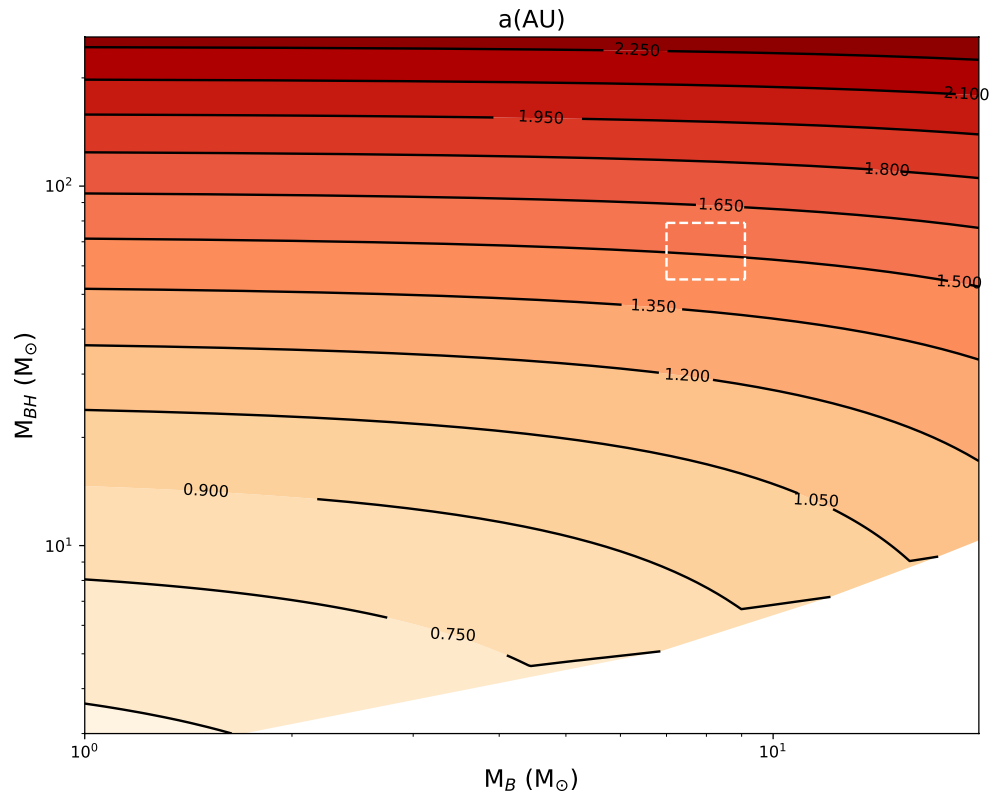


Extended Data Fig. 3 | Variation of $E(B-V)$ with distance in the direction of LB-1. The black circles represent the extinction values corresponding to different distances from the 3D dust map. The green points are the extinction

and distances from SED fitting for each acceptable model of the B star. The red cross marks the extinction value from the 3D dust map at 4.23 kpc, while the red dashed line shows the Gaia distance of 2.14 kpc.

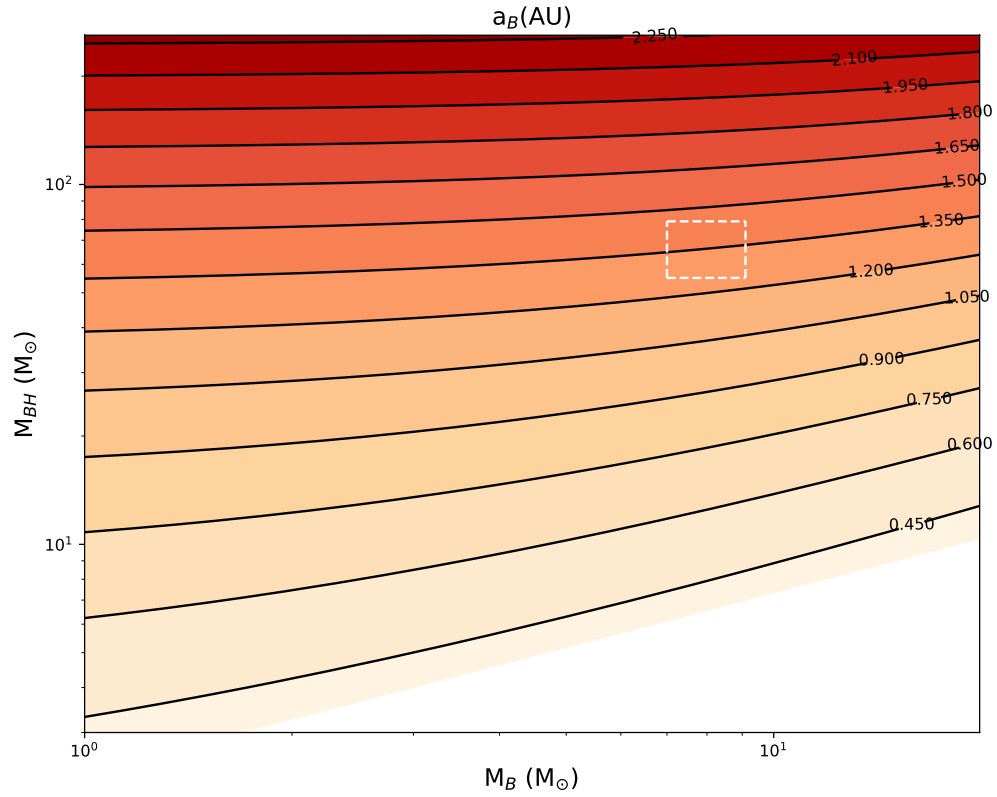


Extended Data Fig. 4 | Search for periodicities for LB-1 with the Lomb–Scargle method. The radial-velocity curve from LAMOST, GTC and Keck observations is being used here. The highest peak corresponds to the orbital period of ~ 78.9 d.



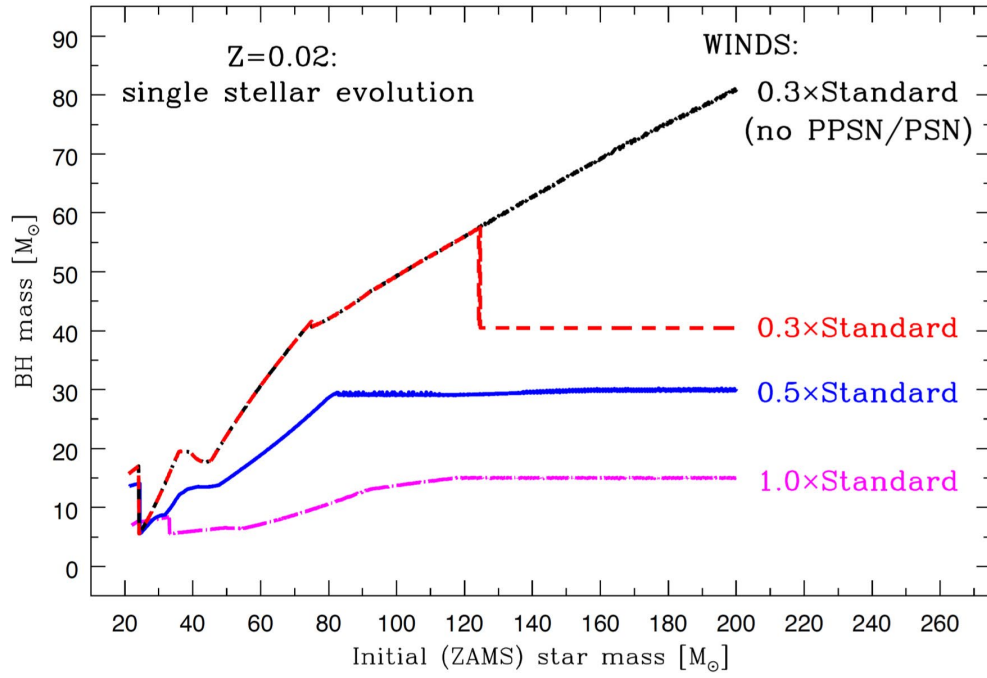
Extended Data Fig. 5 | Separation a as a function of M_B and M_{BH} . Here a is calculated from Kepler's third law for each pair of M_B (B-star mass) and M_{BH} (black-hole mass). The contours and colours both represent the values of a . The

white dashed lines in the contour plot outline a valid region of the separation of the binary system. It comes from the limitations on M_B , $(7-9.1)M_\odot$, and on M_{BH} , $(55-79)M_\odot$. See Methods for details.



Extended Data Fig. 6 | Semi-major axis of the orbit of the B-star a_B as a function of M_B and M_{BH} . Here a_B is calculated from Kepler's third law for each pair of M_B (B-star mass) and M_{BH} (black-hole mass). The contours and colours

both represent the values of a_B . The white dashed lines in the contour plot outline a valid region for the semi-major axis of the B star. It comes from the limitations on M_B , $(7-9.1)M_\odot$, and on M_{BH} , $(55-79)M_\odot$. See Methods for details.



Extended Data Fig. 7 | Black-hole mass versus initial mass in the zero age main sequence (ZAMS) for single stars. For standard wind mass-loss prescriptions, only low-mass black holes are predicted: $M_{\text{BH}} < 15M_{\odot}$ (pink curve). However, for reduced wind mass loss, much heavier black holes are formed: $M_{\text{BH}} = 30M_{\odot}$ for winds reduced to 50% (blue curve) and $M_{\text{BH}} = 60M_{\odot}$ for winds

reduced to 30% (red curve) of the standard values. Note that to reach $M_{\text{BH}} = 80M_{\odot}$ (black curve) it is necessary to switch off pair-instability pulsation supernovae (PPSN) or pair-instability supernovae (PSN), which severely limit black-hole masses.

Extended Data Table 1 | Spectral observations of LB-1

Instrument	Date	Exposure Time (second)	Phase	RV_B (km/s)	RV_α^a (km/s)
(1)	(2)	(3)	(4)	(5)	(6)
LAMOST	2016.11.07	600×15	0.47	39.5±3.4	20.9±4.9
	2016.11.08	600×11	0.48	39.0±3.4	21.0±4.9
	2016.11.23	600×12	0.67	−11.5±3.3	29.8±4.9
	2016.11.26	600×13	0.71	−22.2±3.3	32.7±4.9
	2016.11.28	600×13	0.74	−19.8±3.4	32.1±4.8
	2016.12.01	600×14	0.77	−22.6±3.3	31.8±4.8
	2016.12.02	600×14	0.79	−24.6±3.4	31.5±4.8
	2016.12.05	600×12	0.83	−14.0±3.6	29.6±4.8
	2016.12.06	600×7	0.84	−11.3±3.3	31.5±4.9
	2016.12.17	600×13	0.98	18.0±3.4	28.8±4.9
	2016.12.26	600×9	0.09	58.7±4.9	25.5±5.0
	2017.01.04	600×8	0.20	79.6±3.4	24.2±4.9
	2017.01.05	600×8	0.22	82.3±3.4	24.0±4.9
	2017.01.06	600×7	0.23	81.9±4.0	21.1±4.9
	2017.11.18	600×8	0.24	81.0±3.2	22.4±4.9
	2017.11.19	600×8	0.25	85.2±3.6	21.7±4.9
	2017.11.24	600×10	0.31	78.1±3.4	21.7±4.9
	2017.12.11	600×11	0.53	20.5±3.7	23.7±4.9
	2017.12.17	600×8	0.60	−2.9±3.4	24.9±4.9
	2017.12.21	600×8	0.66	−11.6±3.5	25.7±4.9
	2018.01.16	600×8	0.98	26.7±3.4	24.6±4.9
	2018.01.23	600×9	0.07	49.6±3.8	25.0±4.9
	2018.01.24	600×8	0.09	51.3±3.8	23.9±4.9
	2018.02.12	600×8	0.33	79.1±3.4	22.2±4.9
	2018.02.22	600×7	0.45	45.7±3.3	23.6±4.9
	2018.03.23	600×3	0.82	−22.3±3.4	26.4±5.0
GTC	2017.12.02	V 30×3, R 30×3, I 30×3	0.41	59.5±1.5	24.7±2.8
	2017.12.07	V 30×3, R 30×3, I 30×3	0.47	42.5±1.2	27.1±3.0
	2017.12.10	V 30×3, R 30×3, I 30×3	0.52	26.6±1.3	33.5±2.9
	2017.12.17	V 30×3, R 30×3, I 30×3	0.61	0.1±1.3	33.3±3.5
	2017.12.21	V 30×3, R 30×3, I 30×3	0.66	−18.2±4.1	33.5±3.5
	2017.12.26	V 30×3, R 30×3, I 30×3	0.72	−22.7±1.7	35.4±3.0
	2017.12.31	V 30×3, R 30×3, I 30×3	0.79	−18.0±8.5	37.5±3.4
	2018.01.03	V 30×3, R 30×3, I 30×3	0.83	−17.6±1.4	29.8±3.1
	2018.01.10	V 30×3, R 30×3, I 30×3	0.91	3.7±3.0	32.2±2.9
	2018.01.16	V 30×3, R 30×3, I 30×3	0.99	23.8±3.7	29.4±3.4
	2018.01.20	V 30×3, R 30×3, I 30×3	0.04	42.6±1.2	28.3±3.3
	2018.01.27	V 30×3, R 30×3, I 30×3	0.13	62.5±3.2	28.4±3.2
	2018.01.28	V 30×3, R 30×3, I 30×3	0.14	69.7±4.4	22.2±3.1
	2018.02.15	V 30×3, R 30×3, I 30×3	0.37	72.7±2.1	25.8±3.0
	2018.03.04	V 30×3, R 30×3, I 30×3	0.58	6.1±1.2	30.8±3.3
	2018.03.13	V 30×3, R 30×3, I 30×3	0.70	−19.8±1.7	29.8±2.9
	2018.03.16	V 30×3, R 30×3, I 30×3	0.74	−28.9±1.2	33.9±3.2
	2018.03.24	V 30×6, R 30×6, I 30×6	0.84	−25.2±1.5	29.5±2.9
	2018.03.29	V 30×3, R 30×3, I 30×3	0.90	1.6±1.4	35.4±3.6
	2018.04.07	V 30×3, R 30×3, I 30×3	0.01	29.0±2.8	29.5±3.0
	2018.04.26	V 30×3, R 30×3, I 30×3	0.26	77.4±3.7	22.6±3.0
Keck	2017.12.04	600	0.44	52.8±1.4	26.0±1.5
	2017.12.09	300	0.50	32.7±1.4	26.5±1.4
	2017.12.10	300	0.51	28.2±1.4	31.7±1.5
	2017.12.24	600	0.69	−18.3±1.4	37.2±1.3
	2017.12.29	600	0.75	−22.9±1.4	37.0±1.4
	2017.12.31	500	0.78	−21.8±1.4	36.2±1.3
	2018.01.06	600	0.86	−13.5±1.4	34.5±1.3

See Methods section 'Discovery and follow up observations of LB-1', 'Radial-velocity measurements' for details.

^aThe RV_α corresponds to the 1/3 height method.

Extended Data Table 2 | H α measurement with different methods

Width/method	K_{α} (km/s)	Uncertainty		$V_{0\alpha}$ (km/s)	Uncertainty	
		(90%)	(99%)		(90%)	(99%)
2/3 Height	3.9	0.8	1.2	29.1	0.5	0.9
1/2 Height	4.4	0.7	1.0	28.7	0.5	0.7
1/3 Height	6.4	0.8	1.3	28.9	0.6	1.0
1/4 Height	5.8	1.0	1.5	29.2	0.7	1.1
1/5 Height	6.7	1.0	1.6	29.1	0.8	1.2
120km/s	4.1	0.8	1.2	29.2	0.6	0.9
140km/s	4.8	0.8	1.3	29.0	0.6	0.9
170km/s	5.5	0.8	1.3	29.3	0.6	1.0
200km/s	6.0	0.9	1.4	29.4	0.7	1.1
Bary center (no mask)	1.7	0.9	1.5	29.5	0.7	1.1

See Methods section ‘Radial-velocity measurements’ for details.

Extended Data Table 3 | Orbital parameters of LB-1

Parameter	Value	Uncertainty	
		(90% confidence)	(99% confidence)
(1)	(2)	(3)	(4)
e	0.03	0.01	0.01
K_{B}	52.8	0.7	1.0
$V_{0\text{B}}$	28.7	0.5	0.7
K_{α}^a	6.4	0.8	1.3
$V_{0\alpha}^a$	28.9	0.6	1.0

See Methods section 'Period and orbital parameters' for details.
^aThe $K_{0\alpha}$ and $V_{0\alpha}$ correspond to the 1/3 height method.

A gated quantum dot strongly coupled to an optical microcavity

<https://doi.org/10.1038/s41586-019-1709-y>

Received: 20 December 2018

Accepted: 9 August 2019

Published online: 21 October 2019

Daniel Najer^{1*}, Immo Söllner¹, Pavel Sekatski¹, Vincent Dolique², Matthias C. Löbl¹, Daniel Riedel¹, Rüdiger Schott³, Sebastian Starosielec¹, Sascha R. Valentin³, Andreas D. Wieck³, Nicolas Sangouard¹, Arne Ludwig³ & Richard J. Warburton¹

The strong-coupling regime of cavity quantum electrodynamics (QED) represents the light–matter interaction at the fully quantum level. Adding a single photon shifts the resonance frequencies—a profound nonlinearity. Cavity QED is a test bed for quantum optics^{1–3} and the basis of photon–photon and atom–atom entangling gates^{4,5}. At microwave frequencies, cavity QED has had a transformative effect⁶, enabling qubit readout and qubit couplings in superconducting circuits. At optical frequencies, the gates are potentially much faster; the photons can propagate over long distances and can be easily detected. Following pioneering work on single atoms^{1–3,7}, solid-state implementations using semiconductor quantum dots are emerging^{8–15}. However, miniaturizing semiconductor cavities without introducing charge noise and scattering losses remains a challenge. Here we present a gated, ultralow-loss, frequency-tunable microcavity device. The gates allow both the quantum dot charge and its resonance frequency to be controlled electrically. Furthermore, cavity feeding^{10,11,13–17}, the observation of the bare-cavity mode even at the quantum dot–cavity resonance, is eliminated. Even inside the microcavity, the quantum dot has a linewidth close to the radiative limit. In addition to a very pronounced avoided crossing in the spectral domain, we observe a clear coherent exchange of a single energy quantum between the ‘atom’ (the quantum dot) and the cavity in the time domain (vacuum Rabi oscillations), whereas decoherence arises mainly via the atom and photon loss channels. This coherence is exploited to probe the transitions between the singly and doubly excited photon–atom system using photon-statistics spectroscopy¹⁸. The work establishes a route to the development of semiconductor-based quantum photonics, such as single-photon sources and photon–photon gates.

A resonant, low-loss, low-volume cavity boosts greatly the light–matter interaction so that cavity QED can potentially provide a highly coherent interface between single photons and single atoms. The metric for the coherence is the cooperativity C , the ratio of the coherent coupling squared to the loss rates, $C = 2g^2/(\kappa\gamma)$ (g is the coherent photon–atom coupling, κ is the cavity loss rate and γ is the decay rate of the atom into non-cavity modes). The potential for achieving a high cooperativity gives cavity QED a central role in the development of high-fidelity quantum gates.

In the microwave domain, a transmon ‘atom’ exhibits strong coupling to a cavity photon⁶, facilitating remote transmon–transmon coupling via a virtual photon. Recently, the transmon was replaced with a semiconductor quantum dot (QD), and coupling was observed between a microwave photon and both charge¹⁹ and spin qubits^{20–22}. In contrast to microwave photons, optical-frequency photons can carry quantum information over very large distances and therefore play an indispensable role in quantum communication. Creating an optical photon–photon gate depends critically on a high-cooperativity

photon–atom interface and on efficient photonic engineering⁵. Cavity QED can potentially achieve both simultaneously. Translating these concepts to the solid state is important for developing on-chip quantum technology. The most promising solid-state ‘atom’ is a self-assembled semiconductor QD: an InGaAs QD in a GaAs host is a bright and fast emitter of highly indistinguishable photons^{23,24}, and a QD spin provides the resource required for atom–photon and photon–photon gates. However, a low-noise, high-cooperativity and high-efficiency interface between a single photon and a single QD does not yet exist.

In QD cavity QED, one key problem is the almost ubiquitous observation of scattering from the bare cavity even at the QD–cavity resonance^{10,11,13–17}. This ‘cavity feeding’ is the manifestation of complex noise processes in the semiconductor host¹¹. Another key problem is the trade-off between the coupling g and the loss rates κ and γ in monolithic devices, for instance, in micropillar^{8,23} or photonic-crystal cavities^{9–11,13–15}: as such devices are made smaller, in an attempt to boost g , both κ and γ tend to increase. The increase in κ , reflecting a deterioration in the quality factor (Q) of the microcavity, arises on account

¹Department of Physics, University of Basel, Basel, Switzerland. ²Laboratoire des Matériaux Avancés (LMA), IN2P3/CNRS, Université de Lyon, Lyon, France. ³Lehrstuhl für Angewandte Festkörperphysik, Ruhr-Universität Bochum, Bochum, Germany. *e-mail: daniel.najer@unibas.ch

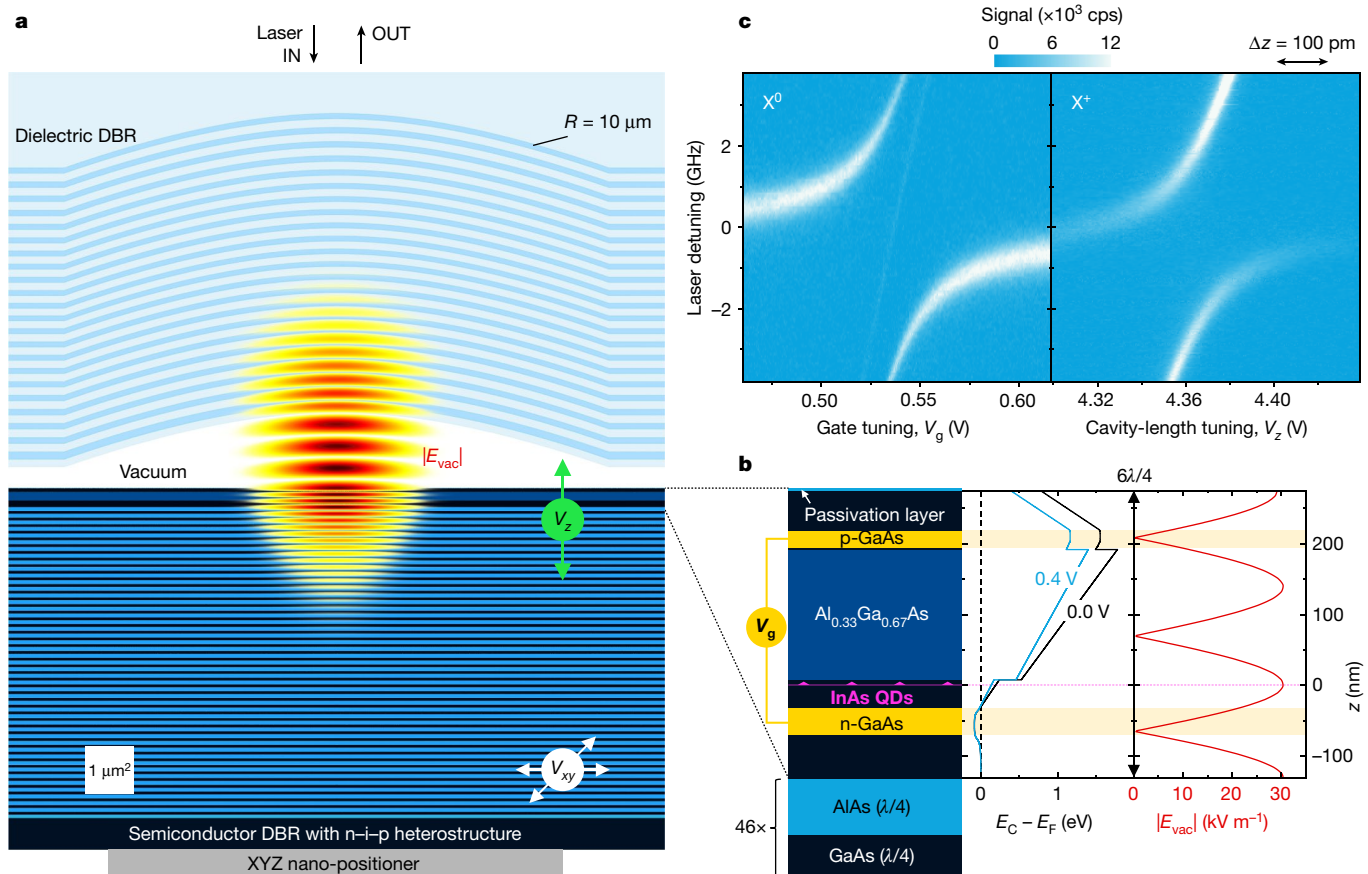


Fig. 1 | Gated QD in a tunable microcavity: design and realization.

a, Simulation of the vacuum electric field $|E_{vac}|$ in the microcavity (image to scale). The bottom mirror is a distributed Bragg reflector (DBR) consisting of 46 pairs of AlAs (thickness $\lambda/4$) and GaAs ($\lambda/4$). (λ is the wavelength in each material.) The top mirror is fabricated in a silica substrate. It has a radius of curvature of $R = 10 \mu\text{m}$ and consists of 22 pairs of silica ($\lambda/4$) and tantala ($\lambda/4$). The layer of QDs is located at the vacuum field anti-node, one wavelength beneath the surface. The vacuum gap has a height of $3\lambda/2$. The voltage V_{xy} (V_z) controls the lateral (vertical) position of the QD with respect to the fixed top mirror. **b**, The top part of the semiconductor heterostructure. A voltage V_g is applied across the n-i-p diode. V_g controls the QD charge via Coulomb blockade and within a Coulomb blockade plateau the exact QD optical

frequency via the d.c. Stark effect. Free-carrier absorption in the p layer²⁸ is minimized by positioning it at a node of the vacuum field. A passivation layer suppresses surface-related absorption²⁵. E_c , conduction band edge; E_F , Fermi energy. **c**, Laser detuning (Δ_L) versus cavity detuning (Δ_C) of a neutral QD exciton (X^0) and a positively charged exciton (X^+) in the same QD. Cavity detuning is achieved by tuning the QD at a fixed microcavity frequency (X^0) and by tuning the microcavity frequency at a fixed QD frequency (X^+). For X^0 , the weak signal close to the bare-exciton frequency arises from weak coupling to the other orthogonally polarized X^0 transition and is unrelated to cavity feeding (see Extended Data Fig. 3a, b). Data in **c** are from QD1 (see Fig. 2) at a magnetic field of $B = 0.00 \text{ T}$.

of increased scattering and absorption; the increase in γ reflects an inhomogeneous broadening in the emitter frequency. The increase in the loss rates is only partly a consequence of fabrication imperfections. An additional factor is the GaAs surface, which pins the Fermi energy mid-gap, resulting in surface-related absorption²⁵ and charge noise.

Here, the QD exhibits close-to-transform-limited linewidths even in the microcavity; the microcavity has an ultrahigh Q factor but small mode volume. Both the frequency and lateral position of the cavity can be tuned in situ via a nano-positioner (Fig. 1a). The QD exciton is far in the strong-coupling regime of cavity QED ($g \gg \kappa, g \gg \gamma$). Strong coupling is achieved on both neutral and charged excitons in the same QD by tuning both the QD charge and the microcavity frequency in situ. The output is close to a simple Gaussian beam. We achieve a cooperativity of $C = 150$, crucially eliminating cavity feeding, and find other sources of noise to be very weak. Equivalently, the β factor, the probability of the excited atom emitting into the cavity mode, is as high as 99.7%. The coherence of the coupled QD-cavity system is demonstrated most clearly by the observation of a very clear atom-photon exchange in the time domain (a vacuum Rabi oscillation).

The design of the QD microcavity is guided by three principles. First and foremost, a self-assembled QD benefits enormously from electrical

control via the conducting gates of a diode structure. A gated QD in high-quality material gives close-to-transform-limited linewidths²⁶, control over both the optical frequency via the Stark effect and the QD charge state via Coulomb blockade²⁷. We therefore include electrical gates in the cavity device. This is non-trivial: the gates themselves, n-doped and p-doped regions in the semiconductor, absorb light via free-carrier absorption, and they are not obviously compatible with a high- Q cavity. Also, the gates inevitably create electric fields in the device, resulting in absorption via the Franz-Keldysh mechanism. Second, in order to achieve narrow QD linewidths in the cavity, we minimize the area of the free GaAs surface to reduce surface-related noise. Finally, we include a mechanism for in situ tuning of the cavity, both in frequency and in lateral position, to allow a full exploration of the parameter space.

We employ a miniaturized Fabry-Pérot cavity consisting of a semiconductor heterostructure and external top mirror (Fig. 1a). The heterostructure has an n-i-p design with the QDs in the intrinsic (i) region (Fig. 1b and Methods). The QDs are located at an antinode of the vacuum field; the p layer is located at the node of the photon field to minimize free-carrier absorption²⁸ (Fig. 1b). Mobile electrons absorb considerably less than mobile holes²⁸, so that it is not imperative to place the n

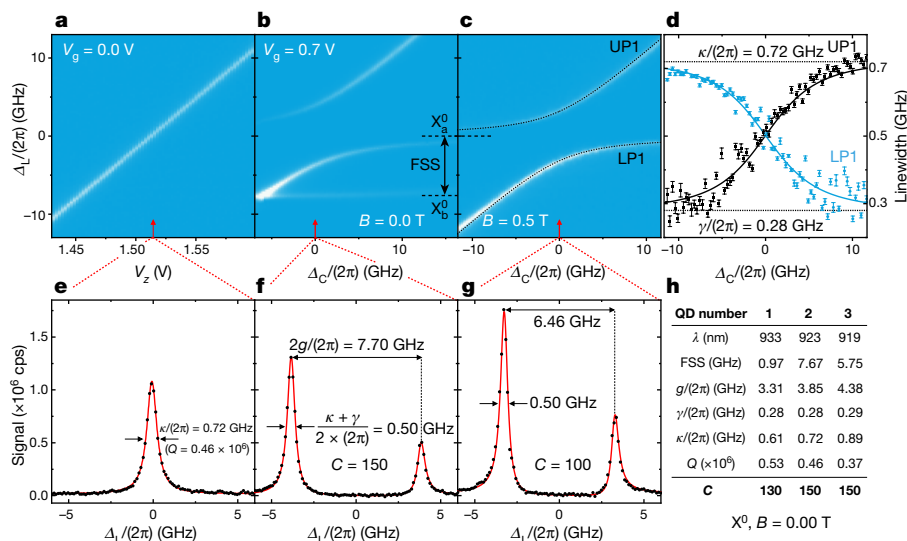


Fig. 2 | Strong coupling of a QD exciton in the microcavity. Spectra recorded by measuring the photons scattered by the microcavity–QD system at a temperature of 4.2 K, rejecting the reflected laser light with a polarization-based dark-field technique^{26,29}. Data shown were taken on the X^0 transition. **a, e**, Signal measured with the QD far-detuned from the microcavity to determine the photon loss rate κ (equivalently, the quality factor Q). **b, f**, X^0 at a magnetic field of $B = 0.00$ T, showing strong coupling to one FSS transition and weak coupling to the other (there is an almost perfect alignment of the X^0 and microcavity axes). From the spectra, we determine g , κ , γ and C (as defined in the main text). **c, d, g**, X^0 at $B = 0.50$ T: the magnetic field increases the FSS. C is

smaller than at $B = 0.00$ T because the X^0 transitions become circularly polarized and couple less strongly to the linearly polarized microcavity mode. The simple avoided crossing in **c** enables the determination of κ and γ by using data at all values of Δ_c . The dotted lines in **c** and solid lines in **d–g** are fits to a solution of the Jaynes–Cummings Hamiltonian in the limit of very small average photon occupation¹⁷. **h**, Summary of strong-coupling parameters recorded for X^0 at $B = 0.00$ T on three separate QDs using the same microcavity mode. In all three cases, $C > 100$. Error bars in **d** are one standard error. Data in **a–g** are from X^0 in QD2.

doping at a node of the vacuum field. The n layer begins 25 nm ‘below’ the QDs, so that they are in tunnel-contact with the Fermi sea in the n layer; that is, the QDs are in the Coulomb blockade regime. The bottom mirror is a semiconductor distributed Bragg reflector (DBR) and the top mirror consists of a 10- μ m-radius dielectric DBR (see Methods). The position of the contacted sample is controlled in situ with respect to the top mirror. We find that surface-related absorption limits the Q factor to 2.0×10^4 . This represented a major problem in the development of this device; to solve it, the GaAs surface was passivated by replacing the native oxide with a few-nanometre-thick alumina layer²⁵. With surface passivation, the fully contacted device had a Q factor close to 10^6 . The mode volume is $1.4\lambda_0^3$ (where λ_0 is the free-space wavelength).

We excite the QD–microcavity system with a resonant laser (continuous-wave), initially with an average photon occupation much less than one ($\langle n \rangle \approx 0.05$), and detect the scattered photons^{26,29}. The fundamental microcavity mode splits into two linearly polarized modes, separated by 32 GHz, predominantly on account of a weak birefringence in the semiconductor. The neutral exciton also splits into a linearly polarized doublet, X_a^0 and X_b^0 , via fine-structure splitting (FSS). QDs are chosen so that the microcavity and X^0 axes are closely aligned. The FSS varies among QDs and can be small enough so that both X_a^0 and X_b^0 couple to the same microcavity mode. In such cases (for example, QD1 in Fig. 2h), this complication can be avoided by applying a magnetic field that pushes X_a^0 and X_b^0 apart via the Zeeman effect. Alternatively, the charged exciton X^+ , which has just one optical resonance at zero magnetic field, can be probed.

When the microcavity and QD optical frequency come into resonance, we observe a clear avoided crossing in the spectral response (Fig. 1c), signifying strong coupling. We achieve strong coupling on different charge states in the same QD (Fig. 1c), also on many different QDs (Fig. 2h and Methods). The cavity–emitter detuning is controlled by tuning either the QD (voltage V_g) or the microcavity (voltage V_z).

At the QD–cavity resonance, mixed states—the polaritons—form. Between the lower and upper polaritons (LP1 and UP1, respectively),

there is no trace of the bare-microcavity mode (Fig. 2f, g). These results demonstrate that cavity feeding has been eliminated, as a consequence of the electrical control via the gates. Coulomb blockade ensures that the QD is always in the charge state that couples to the microcavity mode. (A change of charge detunes the QD, leading to scattering from the bare-microcavity mode.) Phonon-assisted excitation of off-resonant QDs is clearly negligible.

A full spectral analysis determines the parameters g , κ and γ (Fig. 2), giving $\gamma/(2\pi) = 0.28$ GHz (Fig. 2). The transform limit for these QDs is 0.30 ± 0.05 GHz, where the uncertainty accounts for QD-to-QD fluctuations (see Methods). The measured $\gamma/(2\pi)$, 0.28 GHz, corresponds to the ideal limit to within the uncertainties of 10%–20%. The linewidths

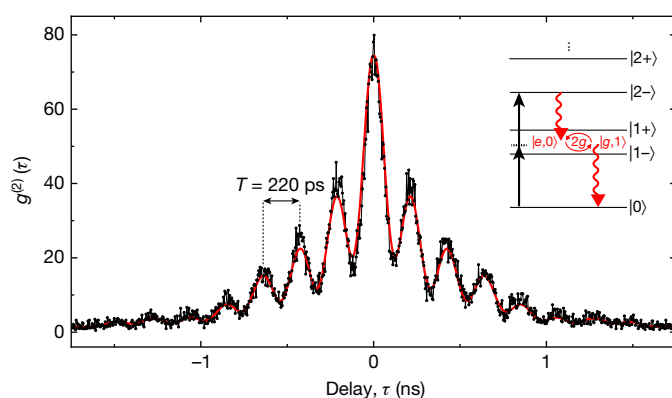


Fig. 3 | Time-resolved vacuum Rabi oscillations. Intensity autocorrelation function $g^{(2)}(\tau)$ as a function of delay τ for $\Delta_c = 0.73g$ (detuned via cavity length) and $\Delta_L = -0.13g$. The inset shows the first few rungs of the Jaynes–Cummings ladder. The laser drives the two-photon transition $|0\rangle \leftrightarrow |2-\rangle$. The solid red line is the result of calculating $g^{(2)}(\tau)$ from the Jaynes–Cummings Hamiltonian using g , κ and γ from the spectroscopy experiments (Fig. 2) and Rabi coupling $\Omega/(2\pi) = 0.16$ GHz. Data are from X^0 in QD1 at $B = 0.40$ T. **|e>**, excited state; **|g>**, ground state.

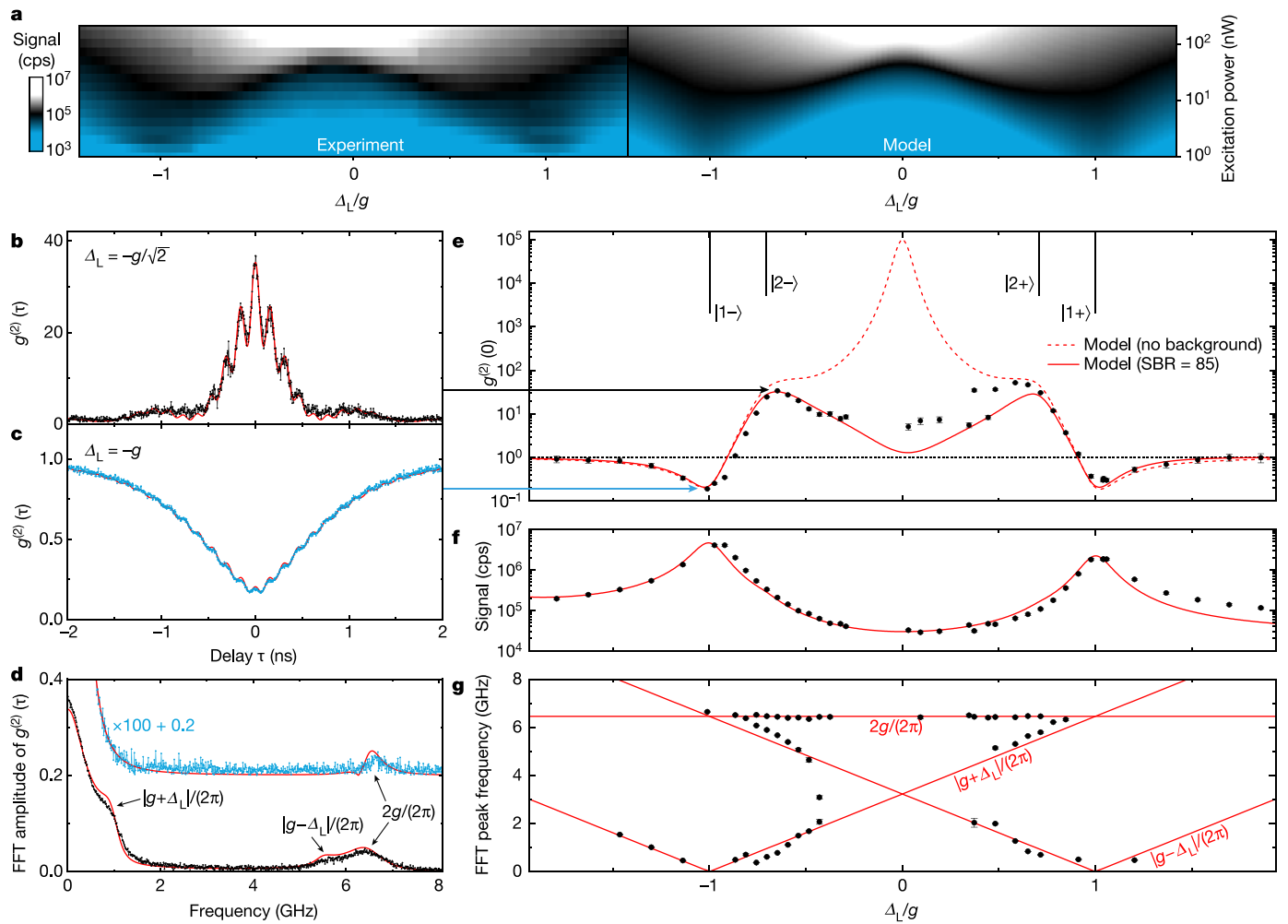


Fig. 4 | Strong coupling versus driving frequency and power. **a**, Signal versus Δ_L for $\Delta_c = 0$. At low power, LP1 and UP1 are clearly observed. As the power increases, the higher rungs of the Jaynes–Cummings ladder are populated. **b**, $g^{(2)}(\tau)$ for $\Delta_c = 0$ and $\Delta_L = -g/\sqrt{2}$. **c**, $g^{(2)}(\tau)$ for $\Delta_c = 0$ and $\Delta_L = -g$. **d**, Fast Fourier transform (FFT) of $g^{(2)}(\tau)$ in **b** and **c**. **e–g**, $g^{(2)}(0)$, signal and FFT peak frequency of $g^{(2)}(\tau)$ versus Δ_L for $\Delta_c = 0$. The solid red lines in **b–g** (‘model’ in **a**) result from a calculation of $g^{(2)}(\tau)$ (signal) from the Jaynes–Cummings Hamiltonian using g , κ

and γ from the spectroscopy experiments (Fig. 2). The Rabi coupling is $\Omega/(2\pi) = 0.14$ – 1.90 GHz (**a**) and 0.07 – 0.11 GHz (**b–g**) and a signal-to-background ratio of SBR = 20 (**a**) and 85 (**b–g**) was included. In **e**, the dashed red line shows the theoretical limit without the laser background. Error bars in **e–g** are one standard error. Data in **a** are from X⁺ in QD1 at $B = 0.00$ T; data in **b–g** are from X⁰ in QD2 at $B = 0.50$ T.

in the microcavity match the best QD linewidths reported so far²⁶. The coupling g lies in the gigahertz regime, pointing to potentially very fast quantum operations. g corresponds closely to that expected considering the geometry of the device (Fig. 1b) and the QD optical dipole. For QD2 at zero magnetic field, $g/\gamma = 14$ and $g/\kappa = 5.3$, corresponding to a cooperativity of $C = 2g^2/(\kappa\gamma) = 150$. Equivalently, the β factor³⁰ is $\beta = 2C/(2C + 1) = 99.7\%$. A high cooperativity is achieved on all QDs within the spectral window of the microcavity (Fig. 2h).

To probe the coherence of the coupled photon–exciton system, we look for a photon–atom exchange, that is, a ‘vacuum Rabi oscillation’^{31,36,31}. We drive the system at a frequency positively detuned from LP1 and record the two-photon autocorrelation $g^{(2)}(\tau)$ (where τ is the delay) without spectral filtering (Fig. 3). Coherent oscillations are observed, and their period, 220 ps, corresponds exactly to 2π divided by the measured frequency splitting of the polaritons at this cavity detuning (see Methods).

These oscillations can be understood in terms of the Jaynes–Cummings ladder (Fig. 3 inset). The laser drives weakly the two-photon transition $|0\rangle \leftrightarrow |2\rangle$. $|2\rangle$ decays by emitting two photons. Detection of the first photon leaves the system in a superposition of the eigenstates $|1\rangle$ and $|1+\rangle$ such that a quantum beat takes place. Detection of

the second photon projects the system into the ground state $|0\rangle$, stopping the quantum beat (Supplementary Information section II). The large $g^{(2)}(0)$ (80 in this experiment) confirms that the photon states with quanta $n \geq 2$ are preferentially scattered^{10,13}.

The measured $g^{(2)}(\tau)$ is fully described by a numerical solution of the Jaynes–Cummings model: the standard Hamiltonian, along with the parameters determined by the spectroscopy experiments (Supplementary Information section I), gives excellent agreement with the experimental result (Fig. 3). The vacuum Rabi oscillations are sensitive to decoherence—not just to the loss processes, but also to pure dephasing of the emitter. Including pure dephasing into the theory improves slightly the quantitative description of $g^{(2)}(\tau)$: the pure dephasing rate is $(10 \pm 2)\%$ of the measured linewidth (Supplementary Information section I.F).

The photon statistics change a lot as a function of both laser detuning, Δ_L , and cavity detuning, Δ_c (both defined with respect to the bare exciton). For $\Delta_c = 0$, $g^{(2)}(0)$ is highly bunched at the two-photon resonance, $\Delta_L = -g/\sqrt{2}$ (Fig. 4b), but highly anti-bunched at the single-photon resonance, $\Delta_L = -g$ (Fig. 4c). The anti-bunching is a demonstration of photon blockade². When driving $|0\rangle \leftrightarrow |1\rangle$, $g^{(2)}(0)$ is limited by the weak two-photon resonance to the $|2\rangle$ state.

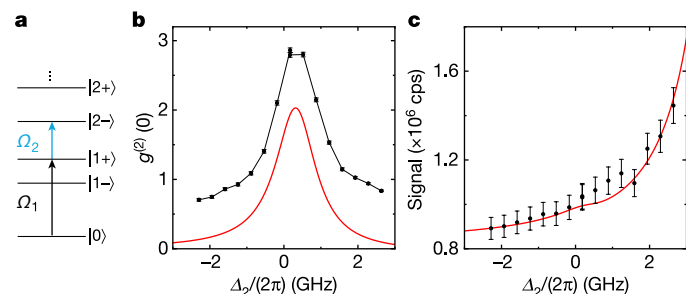


Fig. 5 | Photon-statistics spectroscopy. **a**, Laser 1 is on resonance with the $|0\rangle \leftrightarrow |1+\rangle$ transition (black arrow; detuning $\Delta_1 = 0$); laser 2 is scanned across the $|1+\rangle \leftrightarrow |2-\rangle$ transition (blue arrow; detuning Δ_2). **b**, $g^{(2)}(0)$ versus Δ_2 , showing a pronounced resonance at $\Delta_2 = 3\Delta_C/2 - \Delta_1$. The red solid line is the result of an analytical calculation based on the Jaynes–Cummings Hamiltonian (Supplementary Information section III) with Rabi couplings $\Omega_1/(2\pi) = 0.05$ GHz and $\Omega_2/(2\pi) = 0.45$ GHz. The offset in the experimental data with respect to the theory reflects additional coincidences arising from off-resonant, two-photon absorptions not included in the model. **c**, Signal versus Δ_2 . The signal increases with increasing Δ_2 owing to off-resonant driving of the $|0\rangle \leftrightarrow |1-\rangle$ transition by laser 2. Error bars in **b–c** are one standard error. All data are from X^0 in QD2 at $B = 0.50$ T; $\Delta_C/(2\pi) = 0.31$ GHz, $\Delta_1/(2\pi) = 0.17$ GHz.

This interpretation is confirmed by the weak oscillations in $g^{(2)}(\tau)$ (Fig. 4c), which are established upon the decay of $|2-\rangle$. Further confirmation of this interpretation is provided by QD3, for which g is larger. This increases the detuning of the two-photon transition and thereby weakens it. For QD3, we find a lower value of $g^{(2)}(0)$, $g^{(2)}(0) = 0.09$. The Jaynes–Cummings model reproduces $g^{(2)}(\tau)$ at the photon blockade, both for $g^{(2)}(0)$ and for the fast oscillations.

The full dependence of $g^{(2)}(0)$ on Δ_1 is plotted in Fig. 4e. In principle, $g^{(2)}(0)$ rises to extremely high values as $\Delta_1 \rightarrow 0$. In practice, the scattered signal becomes weaker and weaker as $\Delta_1 \rightarrow 0$, so that $g^{(2)}(0)$ reaches a peak and is then pulled down by the Poissonian statistics of the small leakage of laser light into the detector channel (Fig. 4e). $g^{(2)}(\tau)$ is a rich function: its Fourier transform shows in general three frequencies, corresponding to $2g$ (see Supplementary Information section II.D.3), $|g - \Delta_1|$ and $|g + \Delta_1|$ (Fig. 4d, g). All this complexity is described by the Jaynes–Cummings model, which gives excellent agreement with the experimental $g^{(2)}(\tau)$ in all respects.

As the laser power increases, there is a spectral resonance at the first-to-second-rung transitions, LP2 and UP2, and a strong resonance at $\Delta_1 = 0$ at the highest powers (Fig. 4a). This is also in precise agreement with the Jaynes–Cummings model (Fig. 4a) and reflects the bosonic enhancement of the transitions between the higher-lying rungs of the Jaynes–Cummings ladder. At the highest powers, $\langle n \rangle \approx 1.7$ when driving LP1 or UP1, increasing to $\langle n \rangle \approx 1.6$ when driving at the bare-cavity frequency. This experiment provides an opportunity to measure the quantum efficiency of the system. Given the success of the Jaynes–Cummings model, we can calculate at each laser power the decay rate through the top mirror and hence the expected signal (see Methods). The quantum efficiency of the entire system—that is, from an exciton in the QD to a ‘click’ on the detector—is 8.6%. Importantly, of those photons exiting the top mirror and passing through the dark-field optics, almost all (~94%) make their way into the collection fibre (see Methods). This demonstrates experimentally that the microcavity output is close to a simple Gaussian beam.

In the experiments with a single laser, the second rung of the Jaynes–Cummings ladder is accessed by tuning the laser to a two-photon resonance (Fig. 4b). An alternative is to drive the system with two lasers in a pump–probe scheme. The strong transitions arise from the symmetric-to-symmetric and antisymmetric-to-antisymmetric couplings (for example, $|1-\rangle \leftrightarrow |2-\rangle$ and $|1+\rangle \leftrightarrow |2+\rangle$), which lead to measurable changes in the populations of the states⁶. Here we employ an

alternative, photon-statistics spectroscopy, implementing a theoretical proposal¹⁸, and we present this experiment for the symmetric-to-antisymmetric $|1+\rangle \leftrightarrow |2-\rangle$ transition. The square of the matrix element is just 3% of that associated with the $|1+\rangle \leftrightarrow |2+\rangle$ transition. A pump laser drives the $|0\rangle \leftrightarrow |1+\rangle$ transition on resonance, and a probe laser, which is highly red-detuned with respect to the pump, is scanned in frequency to locate the $|1+\rangle \leftrightarrow |2-\rangle$ transition (Fig. 5a). There is no resonance in the scattered intensity (Fig. 5c), and any resonances lie in the noise (a few per cent). However, there is a clear resonance in $g^{(2)}(0)$ at exactly the expected frequency of $\Delta_2 = 3\Delta_C/2 - \Delta_1$ (Fig. 5b): at the weak $|1+\rangle \leftrightarrow |2-\rangle$ transition the number of scattered photons hardly changes, but there are profound changes in their statistical correlations. Again, the Jaynes–Cummings model describes the experiment well (Fig. 5b, c). Here, a short-time expansion in a truncated Hilbert space (first two rungs of the Jaynes–Cummings ladder) is used to calculate $g^{(2)}(0)$ (Supplementary Information section III).

As an outlook, we offer some perspectives for future development. (a) The device is a potentially excellent single-photon source. For fixed g and γ , the efficiency of photon extraction via the cavity can be maximized by satisfying the condition $\kappa = 2g$ (see Methods). Taking the maximum g reported here, this corresponds to $Q = 3.7 \times 10^4$, which can be achieved with the semiconductor mirror used here and a top mirror with reduced reflectivity. At this relatively low Q , the residual absorption losses in the semiconductor are negligible and, following exciton creation, the efficiency of photon extraction via the top mirror should be as high as 94%. This concept can be profitably combined with lateral excitation, an ‘atom drive’^{3,32} and spin control for the creation of shaped-waveform single photons. (b) The system opens a route towards a photon–photon gate. A key advance here is the coherent exciton–photon interaction, which can be potentially exploited in the Duan–Kimble scheme⁵. Reducing the intrinsic cavity loss by a factor of ten (which is feasible with a more advanced semiconductor design with narrower gates, for instance), the fidelity could be maximized to $F_{pp} = 92\%$ by choosing $\kappa/(2\pi) = 3.8$ GHz (Supplementary Information section IV.D). For both (a) and (b), cavity mode splitting can be eliminated by applying a bias across the semiconductor DBR³³, and a monolithic device could use strain tuning of the QD. On the basis of these considerations, a compact, on-chip, high-cooperativity single-photon–single-QD interface is within reach.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1709-y>.

1. Boca, A. et al. Observation of the vacuum Rabi spectrum for one trapped atom. *Phys. Rev. Lett.* **93**, 233603 (2004).
2. Birnbaum, K. et al. Photon blockade in an optical cavity with one trapped atom. *Nature* **436**, 87–90 (2005).
3. Hamsen, C., Tolazzi, K. N., Wilk, T. & Rempe, G. Two-photon blockade in an atom-driven cavity QED system. *Phys. Rev. Lett.* **118**, 133604 (2017).
4. Zheng, S.-B. & Guo, G.-C. Efficient scheme for two-atom entanglement and quantum information processing in cavity QED. *Phys. Rev. Lett.* **85**, 2392–2395 (2000).
5. Duan, L.-M. & Kimble, H. J. Scalable photonic quantum computation through cavity-assisted interactions. *Phys. Rev. Lett.* **92**, 127902 (2004).
6. Fink, J. M. et al. Climbing the Jaynes–Cummings ladder and observing its \sqrt{n} nonlinearity in a cavity QED system. *Nature* **454**, 315–318 (2008).
7. Kawasaki, A. et al. Geometrically asymmetric optical cavity for strong atom–photon coupling. *Phys. Rev. A (Coll. Park)* **99**, 013437 (2019).
8. Reithmaier, J. et al. Strong coupling in a single quantum dot–semiconductor microcavity system. *Nature* **432**, 197–200 (2004).
9. Yoshie, T. et al. Vacuum Rabi splitting with a single quantum dot in a photonic crystal nanocavity. *Nature* **432**, 200–203 (2004).
10. Faraon, A. et al. Coherent generation of non-classical light on a chip via photon-induced tunnelling and blockade. *Nat. Phys.* **4**, 859–863 (2008).
11. Hennessy, K. et al. Quantum nature of a strongly coupled single quantum dot–cavity system. *Nature* **445**, 896–899 (2007).

12. Rakher, M. T., Stoltz, N. G., Coldren, L. A., Petroff, P. M. & Bouwmeester, D. Externally mode-matched cavity quantum electrodynamics with charge-tunable quantum dots. *Phys. Rev. Lett.* **102**, 097403 (2009).
13. Reinhard, A. et al. Strongly correlated photons on a chip. *Nat. Photon.* **6**, 93–96 (2012).
14. Volz, T. et al. Ultrafast all-optical switching by single photons. *Nat. Photon.* **6**, 605–609 (2012).
15. Ota, Y. et al. Large vacuum Rabi splitting between a single quantum dot and an H0 photonic crystal nanocavity. *Appl. Phys. Lett.* **112**, 093101 (2018).
16. Kuruma, K., Ota, Y., Kakuda, M., Iwamoto, S. & Arakawa, Y. Time-resolved vacuum Rabi oscillations in a quantum-dot–nanocavity system. *Phys. Rev. B* **97**, 235448 (2018).
17. Greuter, L., Starosielec, S., Kuhlmann, A. V. & Warburton, R. J. Towards high-cooperativity strong coupling of a quantum dot in a tunable microcavity. *Phys. Rev. B* **92**, 045302 (2015).
18. Schneebeli, L., Kira, M. & Koch, S. W. Characterization of strong light-matter coupling in semiconductor quantum-dot microcavities via photon-statistics spectroscopy. *Phys. Rev. Lett.* **101**, 097401 (2008).
19. Stockklauser, A. et al. Strong coupling cavity QED with gate-defined double quantum dots enabled by a high impedance resonator. *Phys. Rev. X* **7**, 011030 (2017).
20. Mi, X., Cady, J. V., Zajac, D. M., Deelman, P. W. & Petta, J. R. Strong coupling of a single electron in silicon to a microwave photon. *Science* **355**, 156–158 (2017).
21. Samkharadze, N. et al. Strong spin-photon coupling in silicon. *Science* **359**, 1123–1127 (2018).
22. Landig, A. J. et al. Coherent spin–photon coupling using a resonant exchange qubit. *Nature* **560**, 179–184 (2018).
23. Somaschi, N. et al. Near-optimal single-photon sources in the solid state. *Nat. Photon.* **10**, 340–345 (2016).
24. Ding, X. et al. On-demand single photons with high extraction efficiency and near-unity indistinguishability from a resonantly driven quantum dot in a micropillar. *Phys. Rev. Lett.* **116**, 020401 (2016).
25. Guha, B. et al. Surface-enhanced gallium arsenide photonic resonator with quality factor of 6×10^6 . *Optica* **4**, 218–221 (2017).
26. Kuhlmann, A. V. et al. Charge noise and spin noise in a semiconductor quantum device. *Nat. Phys.* **9**, 570–575 (2013).
27. Högele, A. et al. Voltage-controlled optics of a quantum dot. *Phys. Rev. Lett.* **93**, 217401 (2004).
28. Casey, H. C., Sell, D. D. & Wecht, K. W. Concentration dependence of the absorption coefficient for n- and p-type GaAs between 1.3 and 1.6 eV. *J. Appl. Phys.* **46**, 250–257 (1975).
29. Kuhlmann, A. V. et al. A dark-field microscope for background-free detection of resonance fluorescence from single semiconductor quantum dots operating in a set-and-forget mode. *Rev. Sci. Instrum.* **84**, 073905 (2013).
30. Kuhn, A. & Ljunggren, D. Cavity-based single-photon sources. *Contemp. Phys.* **51**, 289–313 (2010).
31. Kasprzak, J. et al. Up on the Jaynes–Cummings ladder of a quantum-dot/microcavity system. *Nat. Mater.* **9**, 304–308 (2010).
32. Law, C. & Kimble, H. Deterministic generation of a bit-stream of single-photon pulses. *J. Mod. Opt.* **44**, 2067–2074 (1997).
33. Frey, J. A. et al. Electro-optic polarization tuning of microcavities with a single quantum dot. *Opt. Lett.* **43**, 4280–4283 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Design and growth of the heterostructure

The heterostructure is grown by molecular beam epitaxy. It consists of an $n-i-p$ diode with embedded self-assembled InAs QDs grown on top of an AlAs/GaAs DBR with nominal (measured) centre wavelength of 940 nm (920 nm).

The growth on a (100)-oriented GaAs wafer is initiated by a quarter-wave layer (QWL) of an AlAs/GaAs short-period superlattice (18 periods of 2.0-nm-thick GaAs and 2.0-nm-thick AlAs) for stress relief and surface smoothing. The growth continues with 46 pairs of GaAs (67.9 nm) and AlAs (80.6 nm) QWLs forming the ‘bottom’ DBR. The active part of the device consists of a QWL of GaAs (69.8 nm) followed by a 41.0-nm-thick layer of Si-doped GaAs ($n^+, 2 \times 10^{18} \text{ cm}^{-3}$), the back gate. 25.0 nm of undoped GaAs, the tunnel barrier, is subsequently grown, after which InGaAs QDs are self-assembled using the Stranski–Krastanow process and a flushing step³⁴ to blue-shift the QD emission. The layer thicknesses are such that the QDs are located at an antinode of the vacuum electric field. The QDs are capped with an 8.0-nm-thick layer of GaAs. The growth proceeds with an $\text{Al}_{0.33}\text{Ga}_{0.67}\text{As}$ layer (190.4 nm) used as a blocking barrier to reduce the current flow through the diode structure. The heterostructure is completed by 25.0-nm-thick C-doped GaAs (5.0 nm p^+ , $2 \times 10^{18} \text{ cm}^{-3}$ and 20.0 nm p^{++} , $1 \times 10^{19} \text{ cm}^{-3}$), the top gate, and finally a 54.6-nm-thick GaAs capping layer. The heterostructure is shown in Extended Data Fig. 1.

The top gate is centred around a node of the standing wave of the vacuum electric field to minimize free-carrier absorption from the p -doped GaAs. A condition on the tunnel barrier thickness (it is typically less than 40 nm thick to achieve a non-negligible tunnel coupling with the Fermi sea) prevents the back gate from being positioned similarly at a vacuum field node. However, the free-carrier absorption of n^+ -doped GaAs is much smaller than that of p^{++} -doped GaAs at a photon energy of 1.3 eV (absorption coefficient $\alpha \approx 10 \text{ cm}^{-1}$ for n^+ -doped GaAs compared to $\alpha \approx 70 \text{ cm}^{-1}$ for p^{++} -doped GaAs)²⁸. We exploit the weak free-carrier absorption of n^+ -doped GaAs and use a standard 25-nm-thick tunnel barrier. The back gate is thus positioned close to the node of the vacuum electric field but is not centred around the node itself.

Post-growth processing

After growth, individual $2.5 \times 3.0 \text{ mm}^2$ pieces are cleaved from the wafer. The QD density decreases from about 10^{10} cm^{-2} to zero in an approximately centimetre-wide stripe across the wafer. The sample used in these experiments was taken from this stripe and has a density of $7 \times 10^6 \text{ cm}^{-2}$. (The QD density was measured by photoluminescence imaging.)

Separate Ohmic contacts are made to the n^+ and p^{++} layers and a passivation layer is added to the surface. To contact the n^+ layer (the back gate), a local etch in citric acid is used to remove the capping layer, the p^{++} layer, as well as parts of the blocking barrier. NiAuGe is deposited on the new surface by electron-beam physical vapour deposition (EBPVD). Low-resistance contacts to the n^+ layer are formed by thermal annealing. To contact the p^{++} layer (the top gate), another local etch removes the capping layer. A 100-nm-thick Ti/Au contact pad is deposited on the new surface using EBPVD. This contact is not thermally annealed but nevertheless provides a reasonably low-resistance contact to the top gate (Extended Data Fig. 1a).

Following the fabrication of the contacts to the back and top gates, the contacts themselves are covered with photoresist, and the surface of the sample is passivated by chemical treatment. HCl removes a thin oxide layer and a few nanometres of GaAs on the sample surface. After rinsing the sample with deionized water, it is immediately put into an ammonium sulphide ($(\text{NH}_4)_2\text{S}$) bath and subsequently into an atomic layer deposition chamber, where 8 nm of Al_2O_3 is deposited at a temperature of 150 °C. This process is essential for the present device to reduce surface-related absorption, as a high Q factor is achieved only with a surface passivation layer. We can only speculate on the microscopic

explanation of this effect. The passivation procedure reduces the surface density of states, leading to an unpinning of the Fermi energy at the surface. On the one hand, this reduces the Franz–Keldysh absorption in the capping layer; on the other hand, it reduces the absorption from mid-gap surface states. A clear advantage of the surface passivation is that native oxides of GaAs are removed and prevented from re-forming; this not only reduces the probability of surface absorption but also provides a robust and stable termination to the GaAs sample²⁵.

The sample holder contains large Au pads, and Ti/Au and NiAuGe films are connected to the Au pads by wire bonding. Silver paint is used to connect the Au pads to macroscopic wires (twisted pairs).

CO₂ laser ablation of the curved mirror

The template for the curved top mirror is produced by in-house CO₂ laser ablation^{35,36} on a 0.5-mm-thick fused-silica substrate. The radius of curvature of the indentation is 10.5 μm , as measured by confocal scanning microscopy³⁶, and the depth relative to the unprocessed surface is 1.2 μm . After laser ablation, the template is coated with 22 pairs of Ta₂O₅ (refractive index $n=2.09$) and SiO₂ ($n=1.46$) layers (terminating with a layer of SiO₂) by ion-beam sputtering³⁷.

Mirror characterization

Each mirror is characterized by measuring the reflected light intensity at wavelengths outside the stopband. The reflection oscillates as a function of wavelength. We find that these oscillations are a sensitive function of the exact layer thicknesses of the DBR. The transmission is simulated with a one-dimensional transfer matrix calculation, for instance, the Essential Macleod package. A fit is generated, taking the nominal growth parameters as the starting point and making the simplest possible assumption to describe systematic differences between the experimental results and the calculation. In this way, we find that the GaAs (AlAs) layers in the semiconductor DBR start with a physical thickness of 64.6 nm (80.2 nm) for $n=3.49$ ($n=2.92$), reducing linearly to 63.9 nm (79.8 nm). The change arises simply because the growth rate changes slightly during the long process of growing the DBR. Accordingly, we anticipate that the layers in the active layer have actual thicknesses of 38.9 nm (n^+ layer), 29.4 nm (tunnel barrier), 183.3 nm (blocking barrier), 19.0 nm (p^{++} layer), 4.8 nm (p^+ layer) and 55.8 nm (cap). The main consequence of the slight change in growth rate during the growth is that the stopband centre is shifted from 940 nm (design wavelength) to 920 nm. The maximum reflectivity at the stopband centre is not changed substantially by these slight deviations in layer thicknesses.

For technical reasons, the dielectric DBR has a nominal (measured) stopband centre of 1,017 nm (980 nm), that is, it is red-detuned from the QD emission. Because the transmission could not be measured during deposition at a wavelength of 940 nm, a modified quarter-wave stack was chosen, which is expected to have similar transmission (87 ppm) at 1,064 nm and 940 nm. A laser at 1,064 nm was used for in situ characterization. The displacement in the stopband centres between the top and bottom DBRs was an issue only at wavelengths below 915 nm, where the cavity Q factor decreases rapidly with decreasing wavelength. Matching of the two stopband centres would give a high Q factor over a larger spectral range.

Microcavity characterization

A microcavity was constructed using a planar dielectric mirror and the curved dielectric mirror used for the main QD experiment. Both planar and curved silica templates were coated in the same run. With the smallest possible mirror separation of $3\lambda/2$ (limited by the indentation depth of the curved mirror) we determine Q factors of 1.7×10^5 (1.5×10^6) at 920 nm (980 nm) at room temperature. The fundamental microcavity mode splits into a doublet with orthogonal polarizations; at a wavelength of 920 nm, this splitting is typically 13 GHz. These measurements demonstrate the very high quality of the dielectric mirror, in particular the curved dielectric mirror.

The microcavity consisting of the semiconductor mirror and the same curved dielectric mirror has a Q factor of typically 5×10^5 at 920 nm at 4.2 K (Fig. 2), a factor of about 3 larger than the dielectric DBR–dielectric DBR microcavity described above. This increase can be explained by the larger (by a factor of 2) effective cavity length of the semiconductor–dielectric cavity (the group delay of the semiconductor mirror is larger than that of a dielectric mirror owing to the $3\lambda/2$ -thick active layer) and the larger (by a factor of 1.5) finesse. This increase in finesse suggests that at 920 nm the reflectance of the semiconductor mirror is higher than that of the dielectric mirror.

The fundamental mode at a wavelength of 920 nm has a typical polarization splitting of 32 GHz—larger than the polarization splitting of the dielectric DBR–dielectric DBR microcavity (13 GHz at 920 nm). This suggests that the main origin of the polarization splitting is birefringence in the semiconductor induced by strain (AIAs is not exactly lattice-matched to GaAs).

Low-temperature setup and stability

Both the top mirror and the GaAs sample are firmly glued to individual titanium sample holders and mounted inside a titanium ‘cage’ (Extended Data Fig. 1a)³⁸. The holder for the GaAs sample is fixed to a stack of piezo-driven XYZ nano-positioners, whereas the top-mirror holder is fixed to the titanium cage via soft (indium) washers, which act as a flexible material for tilt alignment at room temperature. By observing the cavity with a conventional optical microscope and tightening each screw of the mirror holder individually, Newton rings appear between the two mirrors, which can be centred to ensure mirror parallelism at room temperature. The entire microcavity setup is then inserted in another titanium cage. This outer cage is connected to an optical cage system inside a vacuum tube. The tube is evacuated, flushed with He exchange gas (25 mbar), pre-cooled in liquid nitrogen, and finally transferred into the helium bath cryostat.

To minimise the exposure of the microcavity to acoustic noise, the cryostat is decoupled from floor vibrations via both active and passive isolation platforms (Extended Data Fig. 1b). An acoustic enclosure surrounds both the entire cryostat and the microscope, providing a shield against airborne acoustic noise (Extended Data Fig. 1b). There is no active-feedback mechanism acting on the z -piezo element of the microcavity. Nevertheless, a root-mean-square cavity-length fluctuation³⁶ of about 0.5 pm is measured in the best case, limiting our Q factors to $Q \approx 2.0 \times 10^6$. This corresponds to our highest measured Q factor of $Q = 1.6 \times 10^6$ in the case of a microcavity consisting of the curved top mirror paired with a dielectric bottom mirror of identical coating. This suggests that in the case of the combination of a GaAs sample and a curved dielectric mirror, the Q factor is only slightly reduced by environmental noise.

QD charging

To characterize QD charging, photoluminescence measurements were performed using non-resonant excitation at a wavelength of 830 nm as a function of the voltage applied between the top and bottom gates. Extended Data Fig. 2a shows such a photoluminescence charge map, taken on the sample without the top mirror. Both positive (X^+) and negative (X^-) trions, as well as the neutral exciton (X^0), were identified. The charge states of a QD within the cavity can be recorded in a similar way. To detect all the photoluminescence before filtering by the cavity, a sine wave voltage was applied to the z -piezo element of the cavity so that the cavity was continuously scanned through one free spectral range per integration time window of the spectrometer.

Cross-polarized detection of resonance fluorescence

The behaviour of each QD under resonant excitation can be investigated by suppressing back-reflected laser light in the detection arm and detecting the resonance fluorescence (RF). We achieve this with a dark-field technique²⁹. The optical components are shown in Extended

Data Fig. 1b. The excitation laser passes through a linear polarizer with polarization matched to the reflection of the lower polarizing beam-splitter (PBS). The two PBSs transmit the orthogonal polarization in the vertical direction, the detection channel. The final polarizing element of the excitation channel and the first polarizing element of the detection channel is a quarter-wave plate, which has a dual function. First, by setting the angle of the quarter-wave plate to 45° , the microscope can be operated also in bright-field mode. This is very useful for alignment purposes and for optimization of the out-coupling efficiency. Second, in dark-field mode, the quarter-wave plate allows very small retardances to be introduced, correcting for the slight ellipticity in the excitation polarization state²⁹. The quarter-wave plate allows extremely high bright-field-to-dark-field extinction ratios to be achieved. The microscope can be operated in a set-and-forget mode—once the polarizer and wave plate are aligned, the laser suppression is maintained in the original setup over days²⁹ and even weeks in this case. This very robust operation (despite the fact that control of the wave-plate rotation at the millidegree level is necessary)²⁹ is likely to be a consequence of the effective damping of acoustic and vibrational noise acting on the microscope head in the cavity experiment.

Second-order correlation measurements and single-photon detectors

Second-order correlation measurements are performed with a Hanbury Brown–Twiss (HBT) setup. The signal from the detection fibre (Extended Data Fig. 1b) is sent to a 50:50 fibre beam-splitter and then to two superconducting nanowire single-photon detectors (SNSPDs; Single Quantum Eos). In these experiments, all the photons from the experiment are sent to the HBT setup (no spectral selection is employed). Each SNSPD has a detector efficiency of $\eta_{\text{detector}} \approx 85\%$ and a negligible dark count rate of 10–40 cps. The total timing resolution in the $g^{(2)}$ mode includes the timing resolution of both SNSPDs and the resolution of the time-tagging hardware. In total, it is about 35 ps (full-width at half-maximum), which is well below the vacuum Rabi periods measured in this work.

The dead time of the time-tagging hardware is about 95 ns, which sets a limit for the maximally detectable count rate. To measure count rates higher than about 5×10^6 cps per detector, the 1% arm of the detection fibre is used instead of the 99% arm, and the number of counts is calibrated accordingly.

For the evaluation of $g^{(2)}(\tau)$ we use a time window of 100 ns. For all presented $g^{(2)}(\tau)$ data, we use a bin size of 4 ps. For all presented $g^{(2)}(0)$ values, we perform an FFT of $g^{(2)}(\tau)$ (bin size of 16 ps), then cut all frequency components above 14 GHz and calculate the inverse FFT. In this way, we make sure that the $g^{(2)}(0)$ values are averaged over a time of 35 ps, which is large with respect to the original binning of 16 ps but small with respect to the period of the vacuum Rabi oscillations.

Neutral exciton

An RF scan of QD5 without the top mirror is shown in Extended Data Fig. 2b. In this case, the detuning between the QD and the laser is controlled by fixing the laser frequency and scanning the gate voltage, which detunes the QD resonance frequency via the d.c. Stark shift. Two peaks are observed from the neutral exciton, X^0 . The splitting corresponds to the FSS. By taking several scans for different laser frequencies, a d.c. Stark shift of 240 GHz V^{-1} is determined for this QD. The measured full-width at half-maximum of each neutral exciton peak corresponds to 0.32 GHz, a value close to the typical transform limit of 0.25 GHz for these InGaAs QDs at a wavelength of 940 nm (ref. ³⁹).

Polarization axes

The X^0 polarization axis (hereafter, ‘axis’) varies among QDs. The cavity also has an axis. A complication is that the cavity mode splitting (32 GHz), the X^0 fine structure (1–10 GHz) and the frequency separating the two polaritons in the strong-coupling regime (6–9 GHz) are

all similar. Extended Data Fig. 3a shows an example: full RF scans of the cavity-coupled QD1 are shown, together with their respective line cuts at zero cavity detuning (Extended Data Fig. 3b, f, j). The fundamental cavity mode splits into two modes with linear and orthogonal polarizations. At zero magnetic field ($B = 0.00$ T) the neutral exciton X^0 also splits into two lines with linear and orthogonal polarizations. In the case of QD1 at $B = 0.00$ T, the X^0 and cavity axes are close to parallel, such that one X^0 line couples strongly to one cavity mode and weakly to the other cavity mode, and vice versa for the other X^0 line (Extended Data Fig. 3a). The line cut at one particular cavity frequency shows the polaritons and a weak feature between them (Extended Data Fig. 3b). The analysis including both cavity modes and two X^0 transitions establishes that in Extended Data Fig. 3b the two polaritons arise from strong coupling between one X^0 transition and one cavity mode. The central feature arises from an out-of-resonance response of the strong coupling between the other X^0 transition and the other cavity mode. The bare-cavity mode is not observed at all in the spectral range of Extended Data Fig. 3a.

The QD–cavity couplings in this experiment can be selected in a few ways. First, the X^0 axis varies among QDs. It is not difficult to find a QD with an axis matching closely that of the cavity, so that one X^0 line interacts primarily with one cavity mode and the other X^0 line interacts primarily with the other cavity mode. Extended Data Fig. 3a depicts an example of this behaviour.

Second, application of a small magnetic field pushes the two X^0 lines apart in frequency. At a magnetic field of $B = 0.40$ T, the X^0 lines (QD1) are separated by 12 GHz, so if one X^0 line is resonant with the microcavity, the other X^0 line is far detuned. Extended Data Fig. 3b, f shows an example. In such magnetic fields, the X^0 lines become circularly polarized, so the X^0 axis no longer has a role. The price to pay is a reduction in the coupling parameter g by a factor of $\sqrt{2}$ with respect to the optimal value at zero magnetic field (Extended Data Fig. 3f).

Third, the FSS disappears upon switching to a charged exciton, either X^- or X^+ ; there is just one peak at zero magnetic field (Extended Data Fig. 3i, j), a Zeeman-split doublet at finite magnetic field.

To exploit all three options, we use the power of in situ cavity detuning. When applying a magnetic field or changing the voltage applied to the device, the QD optical frequency changes by many cavity linewidths, but in each case the cavity can be brought into resonance.

Vacuum Rabi frequency versus cavity detuning

Figure 3 shows $g^{(2)}(\tau)$ as a function of delay τ for a cavity detuned by $\Delta_c = 0.73g$ with respect to the emitter. Here we show that vacuum Rabi oscillations in $g^{(2)}(\tau)$ are observed for different values of Δ_c and that the frequency of these oscillations changes according to the change in polariton splitting in the $|1\pm\rangle$ manifold for different values of Δ_c (see Extended Data Fig. 4 and Supplementary Information section V for analytical calculations for the case of $\Delta_c = 0$). The dashed vertical line in Extended Data Fig. 4 depicts the cavity detuning for the data shown in Fig. 3. Consistent with the excellent agreement of the numerical model for $g^{(2)}(\tau)$ with the experiment, an analytical approach to determine the vacuum Rabi period yields $T = 220$ ps, in exact agreement with the experimental observations.

$g^{(2)}(0)$ versus laser and cavity detuning

In the experiment, three frequencies can be tuned in situ: the laser frequency ω_L , the emitter frequency ω_x (via the gate voltage) and the cavity frequency ω_c (via tuning of the cavity length).

Figure 4e shows $g^{(2)}(0)$ as a function of laser detuning Δ_L for a cavity detuning of $\Delta_c = 0$ on QD2 at $B = 0.50$ T. $g^{(2)}(0)$ can be described well by the model and a small laser background. This point is investigated also in other cases. Extended Data Fig. 3c, g, h shows more $g^{(2)}(0)$ measurements of the neutral exciton of QD1 at $B = 0.00$ T and 0.40 T, with close-to-zero cavity detuning (Extended Data Fig. 3c, g) and a cavity detuning of $\Delta_c \approx g$ (Extended Data Fig. 3h).

The in situ tunability of the microcavity can be exploited in an alternative experiment, in which the cavity is detuned and the polaritons are driven resonantly at each cavity detuning. Extended Data Fig. 5a, b shows the behaviour of the first-rung polaritons (LP1 in black, UP1 in red) as a function of Δ_c . Also in this case, the model reproduces the experimental results well. The reason for the slight discrepancy in the $g^{(2)}(0)$ values of the lower polariton at large and negative Δ_c is the fact that the laser starts driving the second fine-structure level, which is weakly coupled to the same cavity mode. This increases slightly the number of single photons in the detection signal, as evidenced by the slight anti-bunching in the experimental data.

Power dependence

The experiments in Figs. 1–5, Extended Data Figs. 3a–c, e–j, 5a, b are all recorded with a weak driving laser, that is, with a mean photon number in the cavity well below 1. We present here the observed behaviour of the system as the power of the driving laser increases.

In Extended Data Fig. 5c we plot the measured and calculated scattering signal measured when driving LP1 (black) and UP1 (red) with increasing excitation power. A striking feature is that the system does not saturate (Extended Data Fig. 5c). This is evidence that the full ladder of Jaynes–Cummings levels exists. To model this power dependence, it is necessary to determine the connection between the Rabi frequency Ω , the input parameter to the model, and the laser power P , the control parameter in the experiment. Clearly, $\Omega \propto \sqrt{P}$. At the lowest powers, only the zeroth and first rungs of the Jaynes–Cummings ladder are populated, so that the $|0\rangle \leftrightarrow |1-\rangle$ and $|0\rangle \leftrightarrow |1+\rangle$ transitions behave like two-level systems: the scattered signal increases linearly with laser power, as expected (Extended Data Fig. 5c).

We parameterize the link between Ω and P by adopting the link for a two-level system, namely $\Omega = \sqrt{\frac{P}{P_0} \frac{\kappa + \gamma}{2} \frac{1}{\sqrt{2}}}$, where P is the laser power (monitored at the 50:50 fibre beam-splitter) and P_0 is a reference power. The signal S is equal to the steady-state photon occupation in the cavity multiplied by the cavity loss rate (κ) and the cavity-to-detector system efficiency (η_{system}), $S = \eta_{\text{system}} \kappa \langle n \rangle$. We calculate $\langle n \rangle$ from the Jaynes–Cummings model with the parameters g , κ and γ determined from the spectroscopy experiment and $(\Delta_c, \Delta_L) = (0, \pm g)$.

The nonlinear power dependence (Extended Data Fig. 5c) enables both P_0 and η_{system} to be determined. A fit to the experimental data leads to $P_0 = 214$ nW ($P_0 = 529$ nW) for LP1 (UP1) and $\eta_{\text{system}} = 12\%$. The difference in P_0 for LP1 and UP1 results in unequal polariton populations at constant input powers, as seen in Fig. 2f, g. The difference in P_0 values probably arises from a polarization-dependent chromaticity in the throughput of the excitation channel of the microscope. The same model gives excellent agreement with the experimental $g^{(2)}(0)$ for both LP1 and UP1 (Extended Data Fig. 5d, e).

The behaviour of the system as a function of driving power can also be explored by measuring the Δ_L dependence of the scattered intensity for $\Delta_c = 0$. Extended Data Fig. 3d, k shows power-dependent RF scans acquired when the bare exciton and cavity are resonant. At low power, LP1 and UP1 are clearly resolved. At higher power, bumps appear at the two-photon LP2 and UP2 resonances. In Extended Data Fig. 3k, there is no resonance close to the bare-cavity mode at low power, enabling us to explore fully the behaviour of the system, even at very large driving powers. At the highest powers, the response is dominated by a feature at $\Delta_L \approx 0$ (Extended Data Fig. 3k). This too is evidence that the full Jaynes–Cummings ladder can be accessed. At the highest powers, the system ‘climbs’ the Jaynes–Cummings ladder because of the bosonic enhancement of photons, so that the average photon occupation is large and the polariton resonances become closer in frequency to the bare-cavity mode. This power dependence can also be described by the model, and very good agreement is found between our numerical model and the data in Extended Data Fig. 3k. (Owing to the presence of

the second fine-structure level in Extended Data Fig. 3d, our numerical model is incomplete in this case.)

Quantum efficiency

The dependence of the scattered intensity on laser power enables us to determine $\eta_{\text{system}} = 12\%$. One contribution to η_{system} is the out-coupling efficiency²³, which is defined as the fraction of photons in the κ channel leaving through the top mirror (rate κ_{top}):

$$\eta_{\text{out}} = \frac{\kappa_{\text{top}}}{\kappa} \times \frac{T_{\text{top}}}{T_{\text{top}} + T_{\text{bottom}} + A} \quad (1)$$

Using the model of the mirrors, we determine (T_{top} , T_{bottom} , A) = (116, 1, 373) ppm at wavelength $\lambda = 923$ nm. Here, T_{top} (T_{bottom}) and A are the fractional intensity losses per round trip via transmission through the top (bottom) mirror and absorption/scattering losses, respectively. This gives $\eta_{\text{out}} = 24\%$.

The system efficiency can be described using a number of additional factors. If the cavity and microscope axes lie at an angle of $\phi = 45^\circ$ to each other, $\eta_{\text{dark-field}} = 50\%$. This is not exactly the case in practice. For X^0 in QD2 ($B = 0.50$ T) it is $\phi = 37^\circ \pm 6^\circ$, resulting in $\eta_{\text{dark-field}} = (63 \pm 10)\%$. Once a photon has entered the detection channel after the dark-field polarization optics, it is coupled into the collection fibre with probability η_{fibre} . Overall,

$$\eta_{\text{system}} = \eta_{\text{out}} \times \eta_{\text{dark-field}} \times \eta_{\text{fibre}} \times \eta_{\text{detector}} \quad (2)$$

The detector has a quantum efficiency of $\eta_{\text{detector}} = 85\%$. From these results, we find that $\eta_{\text{fibre}} = (94^{+6}_{-15})\%$.

The collection fibre is a single-mode optical fibre and supports a propagating Gaussian mode. The high value of η_{fibre} is only achievable with excellent mode matching between the cavity output and the optical fibre, which constitutes experimental proof that the cavity output is described extremely well by a Gaussian mode.

We note that the exciton-to-photon quantum efficiency (the probability of an exciton producing a photon that exits via the κ channel) of the microcavity⁴⁰ is

$$\eta_{\text{cavity}} = \beta \frac{\kappa}{\kappa + \gamma} \quad (3)$$

which is 72% for QD2 ($B = 0.50$ T). By maximizing η_{cavity} for fixed g and γ (by choosing $\kappa = 2g$), the collection efficiency $\eta_{\text{cavity}} \times \eta_{\text{out}}$ into the first lens of the optical setup^{23,24,41,42} can be as high as 94% with the

present microcavity device. The overall exciton-to-detector quantum efficiency is

$$\eta_{\text{exciton}} = \eta_{\text{cavity}} \times \eta_{\text{system}} \quad (4)$$

which is 8.6% in the present experiment.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

34. Wasilewski, Z., Fafard, S. & McCaffrey, J. Size and shape engineering of vertically stacked self-assembled quantum dots. *J. Cryst. Growth* **201–202**, 1131–1135 (1999).
35. Hunger, D., Deutsch, C., Barbour, R. J., Warburton, R. J. & Reichel, J. Laser micro-fabrication of concave, low-roughness features in silica. *AIP Adv.* **2**, 012119 (2012).
36. Greuter, L. et al. A small mode volume tunable microcavity: development and characterization. *Appl. Phys. Lett.* **105**, 121105 (2014).
37. The VIRGO Collaboration. The VIRGO large mirrors: a challenge for low loss coatings. *Class. Quantum Gravity* **21**, S935–S945 (2004).
38. Barbour, R. J. et al. A tunable microcavity. *J. Appl. Phys.* **110**, 053107 (2011).
39. Dalgarno, P. A. et al. Coulomb interactions in single charged self-assembled quantum dots: radiative lifetime and recombination energy. *Phys. Rev. B* **77**, 245311 (2008).
40. Cui, G. & Raymer, M. G. Quantum efficiency of single-photon sources in the cavity-QED strong-coupling regime. *Opt. Express* **13**, 9660–9665 (2005).
41. Wang, H. et al. On-demand semiconductor source of entangled photons which simultaneously has high fidelity, efficiency, and indistinguishability. *Phys. Rev. Lett.* **122**, 113602 (2019).
42. Liu, J. et al. A solid-state source of strongly entangled photon pairs with high brightness and indistinguishability. *Nat. Nanotechnol.* **14**, 586–593 (2019).

Acknowledgements We thank I. Favero for inspiration on surface passivation, H. Thyrestrup Nielsen for support in evaluating $g^{(2)}(\tau)$ with very small binning times, S. Martin for engineering the microcavity hardware, and M. Ho, A. Javadi, P. Lodahl and P. Treutlein for discussions. We acknowledge financial support from SNF projects 200020 156637 and PPOOP2 179109, NCCR QSIT and EPPIC (747866). S.R.V., R.S., A.L. and A.D.W. acknowledge support from BMBF (Q.Link.X 16KIS0867) and DFG (LU2051/1-1).

Author contributions D.N. carried out the surface passivation, the experiments and the detailed data analysis. I.S. developed the numerical model and used it to guide the experiments. P.S. and N.S. developed the analytical theory. M.C.L. contributed to the QD characterization, the dark-field setup and the $g^{(2)}(\tau)$ measurements. S.S. and D.R. developed the microcavity experiment and assisted D.N. in its operation. A.L., D.N., S.S. and R.J.W. designed the heterostructure. S.R.V., R.S., A.D.W. and A.L. fabricated the device for the experiments (molecular beam epitaxy of the heterostructure, post-growth processing of the diode structure). V.D. applied the dielectric coating to the curved-mirror template fabricated by S.S. D.N. and R.J.W. wrote the manuscript with input from all authors.

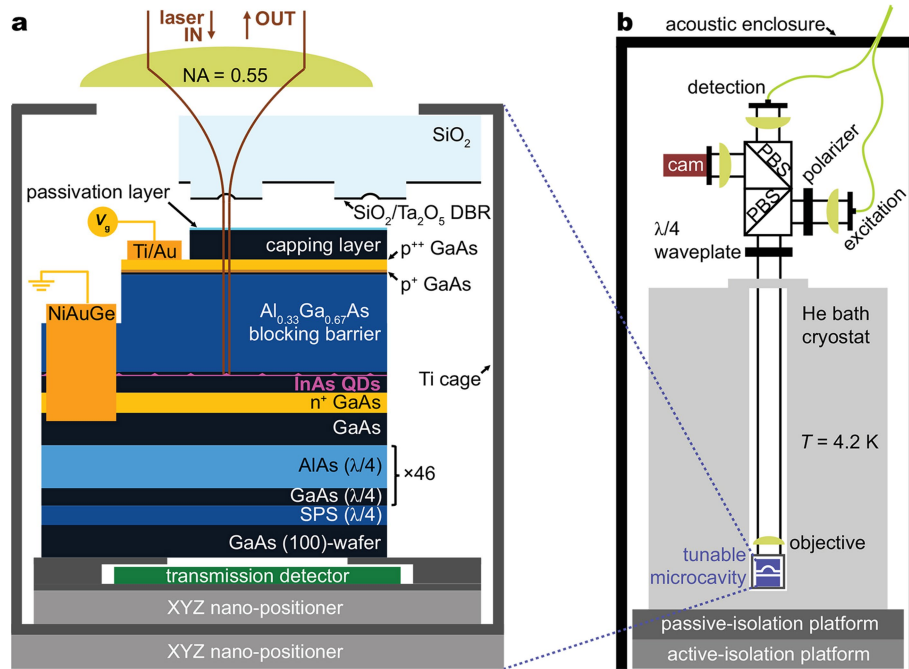
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1709-y>.

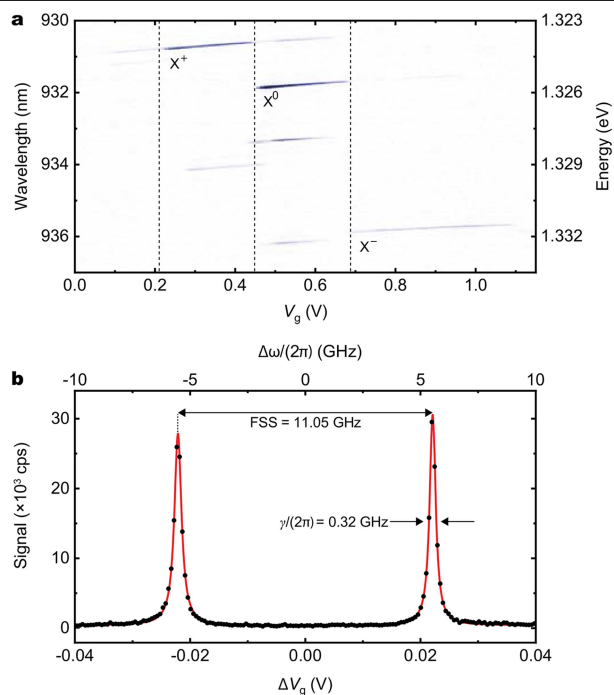
Correspondence and requests for materials should be addressed to D.N.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



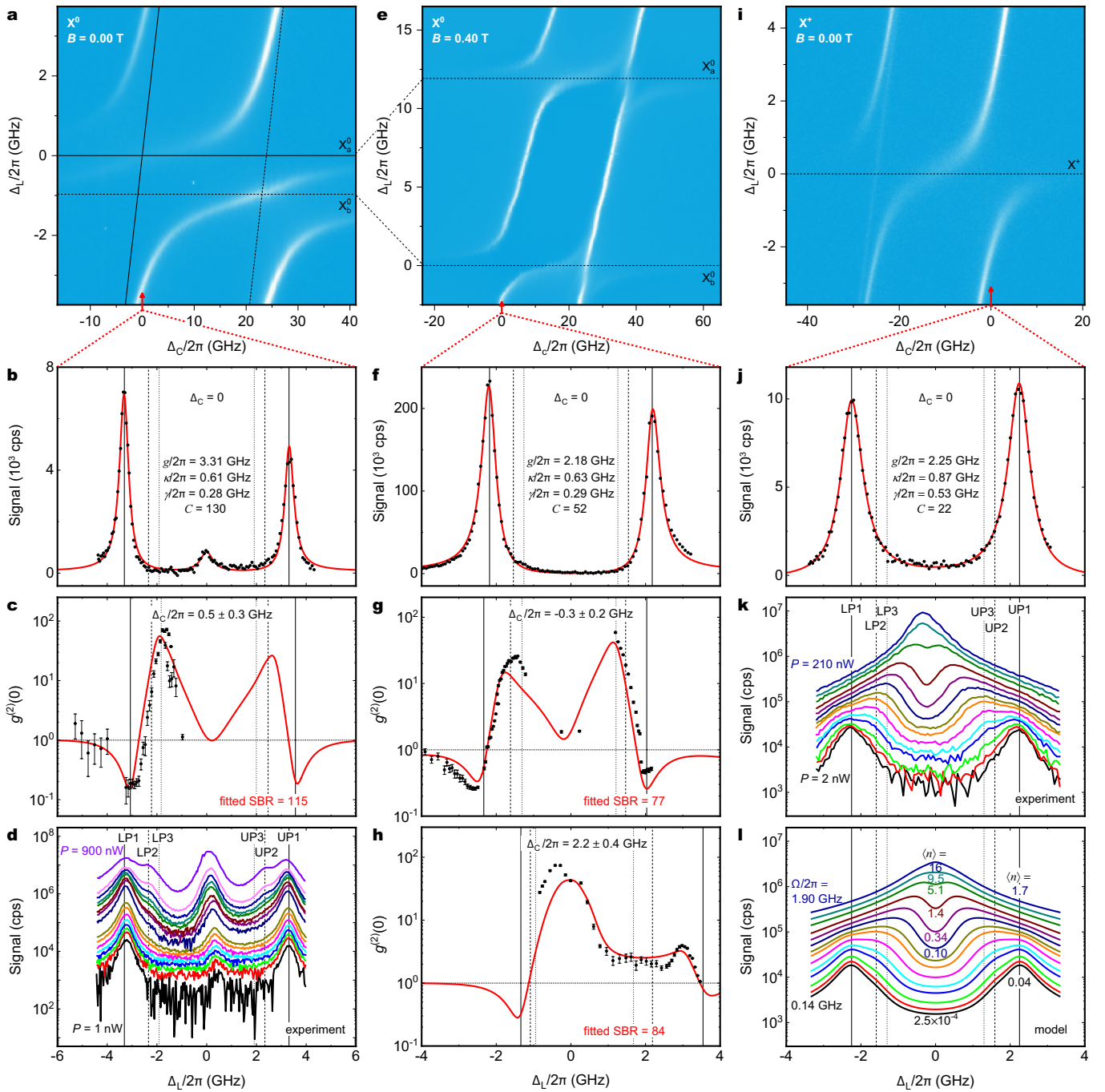
Extended Data Fig. 1 | Tunable-microcavity setup. **a**, The top mirror is fixed to the upper inner surface of a titanium ‘cage’. The sample is mounted on a piezo-driven XYZ nano-positioner, which is fixed to the bottom inner surface of the cage. The nano-positioner allows full in situ spatial and spectral tuning of the microcavity at cryogenic temperatures. The titanium cage resides on another XYZ nano-positioner, enabling close-to-perfect mode matching of the cavity mode to the external laser beam³⁶. **b**, An outer Ti cage, containing the inner Ti cage and a second nano-positioner, is fixed to an optical rod system, which is inserted into a vacuum tube filled with He exchange gas. The optical elements depicted in the image (objective lens, quarter-wave plate, two polarizing beam-splitters (PBSs), polarizer, CMOS camera, two fibre couplers) make up the dark-

field microscope for near-background-free detection of resonance fluorescence²⁹. The back-reflected laser is suppressed by a factor of up to 10^8 by choosing orthogonal polarization states for the excitation and detection channels²⁹. The optical fibre attached to the excitation (detection) arm of the microscope includes a 50:50 (99:1) fibre beam-splitter to monitor the laser power sent into the microscope (reflected from the sample). The cryostat sits on both active- and passive-isolation platforms and is surrounded by an acoustic enclosure to minimize acoustic noise. Both images are schematic representations and are not to scale. The exact layer thicknesses and doping concentrations are given in the text.



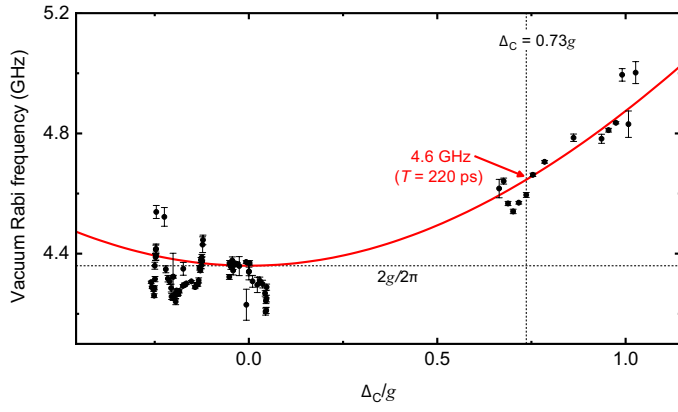
Extended Data Fig. 2 | QD charging and neutral exciton linewidth.

a. Measured photoluminescence signal of non-resonantly excited QD4 ($\lambda = 830$ nm, $P = 200$ nW, $B = 0.00$ T) as a function of gate voltage. The three main charge states of the QD are the positive trion (X^+), neutral exciton (X^0) and negative trion (X^-). Dark blue, maximum number of counts; white, minimum number of counts. **b.** Resonance fluorescence on QD5 (X^0 , $\lambda = 939$ nm, $B = 0.00$ T) excited well below saturation (red solid line, Lorentzian fit). From the measured Stark shift of 240 GHz V^{-1} , a linewidth of 0.32 GHz is obtained, which is close to the typical transform limit of 0.25 GHz for these InGaAs QDs at a wavelength of 940 nm (ref. ³⁹). The splitting arises from the fine structure of X^0 and is 11.05 GHz for QD5.

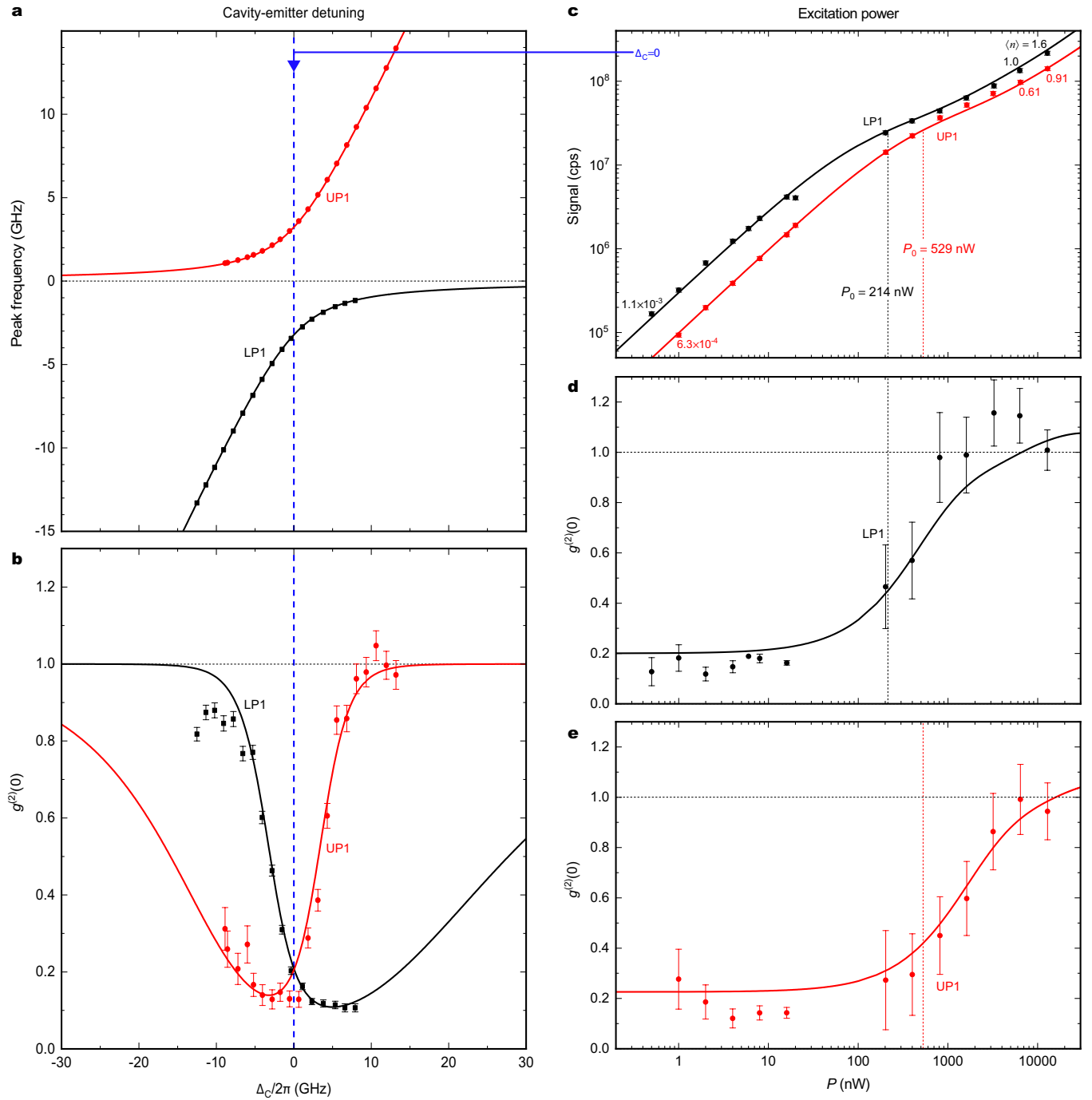


Extended Data Fig. 3 | Spectroscopy on cavity-coupled QD1. **a**, X^0 at $B = 0.00$ T. RF scan revealing two transverse-electromagnetic (TEM_{00}) cavity modes with a polarization splitting of 25 GHz (inclined lines) coupled to two FSS levels of X^0 with a splitting of 1 GHz (horizontal lines). **b**, Line cut at resonance with the ‘left’ cavity mode (as indicated by red arrow). The main peaks arise from coupling of the ‘high’-frequency X^0 transition to one cavity mode; the peak at $\Delta_L = 0$ arises from coupling of the ‘low’-frequency X^0 transition to the same cavity mode. **c**, $g^{(2)}(0)$ versus laser detuning for a cavity detuning close to zero. **d**, Power dependence at resonance. Excitation of the second rung of the Jaynes–Cummings ladder (LP2, UP2) is evident at high powers, as indicated by the dashed vertical lines. **e**, X^0 at $B = 0.40$ T. RF scan revealing that the same TEM_{00} cavity modes couple to the two X^0 transitions. The X^0 transitions are now separated by Zeeman splitting. **f**, Line cut at resonance with the ‘left’ cavity mode. **g**, **h**, $g^{(2)}(0)$ versus laser detuning for two

different cavity detunings: one close to zero and one close to g . **i**, X^+ at $B = 0.00$ T. RF scan of the X^+ transition. **j**, Line cut at resonance with the ‘right’ cavity mode. **k**, **l**, Experimental (**k**) and theoretical (**l**) power dependence at resonance. The excitation of higher rungs of the Jaynes–Cummings ladder is evident by the convergence from the two first-rung polaritons towards the bare-cavity mode with increasing power, leading to a calculated mean photon number in the cavity of up to $\langle n \rangle = 16$. The Hilbert space in the model is truncated to 35 rungs of the Jaynes–Cummings ladder. The slight frequency shift of the signal peak in **k** at maximum laser power is due to an unintended drift of the cavity length during this experiment. In all figures, the vertical lines depict the resonance frequencies for the first three rungs of the Jaynes–Cummings ladder (LP1, UP1: solid; LP2, UP2: dashed; LP3, UP3: dotted) at a particular cavity detuning. Error bars in **c**, **g**, **h** are one standard error.



Extended Data Fig. 4 | Vacuum Rabi frequency versus Δ_c . The data points correspond to measured vacuum Rabi frequencies (determined via FFT of $g^{(2)}(\tau)$) for different cavity detunings Δ_c . The red solid line is the result of an analytical calculation of the polariton splitting in the $|1\pm\rangle$ manifold for different values of Δ_c (see equation (15) in Supplementary Information section III) using a coupling strength measured spectroscopically (Extended Data Fig. 3f). Error bars are one standard error. Data are from X^0 in QD1 at $B = 0.40$ T.



Extended Data Fig. 5 | Spectroscopy of cavity-coupled QD2. **a**, Experimental and theoretical dispersion of the lower (LP1) and the upper (UP1) polariton. **b**, Corresponding experimental and theoretical $g^{(2)}(0)$ values. **c**, Intensity of scattered light from LP1 and UP1 at zero cavity detuning as a function of resonance excitation power. The absence of saturation is due to the population of higher rungs of the Jaynes–Cummings ladder with increasing power. The behaviour at low powers allows the dependence of the Rabi frequency Ω on

laser power P to be determined. This behaviour is parameterized with power P_0 (see text for definition): $P_0 = 214$ nW for LP1 and $P_0 = 529$ nW for UP1 (black and red dashed vertical lines, respectively). The mean photon number $\langle n \rangle$ is shown. **d**, **e**, Corresponding experimental and theoretical $g^{(2)}(0)$ values for LP1 (**d**) and UP1 (**e**). Error bars in **b–e** are one standard error. All data are from X⁰ in QD2 at $B = 0.50$ T.

Imaging work and dissipation in the quantum Hall state in graphene

<https://doi.org/10.1038/s41586-019-1704-3>

Received: 6 May 2019

Accepted: 8 August 2019

Published online: 21 October 2019

There are amendments to this paper

A. Marguerite^{1,4}, J. Birkbeck^{2,4}, A. Aharon-Steinberg^{1,4}, D. Halbertal^{1,3}, K. Bagani¹, I. Marcus¹, Y. Myasoedov¹, A. K. Geim², D. J. Perello^{2*} & E. Zeldov^{1*}

Topology is a powerful recent concept asserting that quantum states could be globally protected against local perturbations^{1,2}. Dissipationless topologically protected states are therefore of major fundamental interest as well as of practical importance in metrology and quantum information technology. Although topological protection can be robust theoretically, in realistic devices it is often susceptible to various dissipative mechanisms, which are difficult to study directly because of their microscopic origins. Here we use scanning nanothermometry³ to visualize and investigate the microscopic mechanisms that undermine dissipationless transport in the quantum Hall state in graphene. Simultaneous nanoscale thermal and scanning gate microscopy shows that the dissipation is governed by crosstalk between counterpropagating pairs of downstream and upstream channels that appear at graphene boundaries as a result of edge reconstruction. Instead of local Joule heating, however, the dissipation mechanism comprises two distinct and spatially separated processes. The work-generating process that we image directly, which involves elastic tunnelling of charge carriers between the quantum channels, determines the transport properties but does not generate local heat. By contrast, the heat and entropy generation process—which we visualize independently—occurs nonlocally upon resonant inelastic scattering from single atomic defects at graphene edges, and does not affect transport. Our findings provide an insight into the mechanisms that conceal the true topological protection, and suggest routes towards engineering more robust quantum states for device applications.

Major progress has been made in recent years in identifying new topological states of matter^{1,2}; however, the extent to which topological protection is manifested in realistic systems and the microscopic mechanisms that lead to its apparent breakdown remain poorly understood. The quantum Hall effect is a prime example of a topologically protected state that exhibits quantized dissipationless electron transport. Although an extremely high degree of conductance quantization has been achieved in engineered systems in gallium arsenide (GaAs) heterostructures and in graphene⁴, quantum Hall devices commonly exhibit small but fundamentally important deviations from the ideal quantized conductance. Various mechanisms that undermine the topological protection have been explored, including imperfect contacts⁵, current-induced breakdown⁴, absence of edge equilibration⁶ and edge reconstruction^{7,8}. Nonetheless, exactly how the dissipation in the quantum Hall regime occurs on a microscopic level has not been directly identified. Here we provide nanoscale imaging of the dissipation processes in the quantum Hall state in graphene and reveal the intricate mechanisms that compromise the apparent global topological protection.

A superconducting quantum interference device—SQUID-on-tip (SOT)⁹, which acts as a nanothermometer (tSOT) with microkelvin

sensitivity³ and has an effective diameter of around 50 nm—was scanned approximately 50 nm above the surface of high-mobility hexagonal boron nitride (hBN)-encapsulated graphene devices (see Methods) at $T = 4.2$ K. Three modalities were used simultaneously (Fig. 1a, see Methods): (i) d.c. thermal imaging, which maps the local temperature variations $T_{dc}(\mathbf{r})$ induced by an externally applied current I_{dc} . The current was chopped at around 94 Hz and $T_{dc}(\mathbf{r})$ was recorded using a lock-in amplifier. (ii) a.c. thermal imaging, in which the tSOT is mounted on a quartz tuning fork and vibrates parallel to the sample surface at a frequency of around 35 kHz with an amplitude x_{ac} of around 8 nm. The resulting $T_{ac}(\mathbf{r}) \approx x_{ac} \partial T_{dc}(\mathbf{r}) / \partial x$ provides a high-sensitivity map of the local temperature gradients³. (iii) Scanning gate mode¹⁰, in which a voltage V_{tg} is applied to the tip and the induced variations in the two-probe, $R_{2p}(\mathbf{r})$, or four-probe, $R_{xx}(\mathbf{r})$, sample resistance is imaged.

Topological protection in an idealized integer quantum Hall system is manifested by three guiding principles⁶. First, in the quantum Hall plateau the current flows along ballistic chiral edge channels with no backscattering and no dissipation except at the current contacts. Second, in the plateau transition regions, dissipation sets in through the bulk of the system. Third, the topological state and the

¹Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot, Israel. ²National Graphene Institute and School of Physics and Astronomy, The University of Manchester, Manchester, UK. ³Present address: Department of Physics, Columbia University, New York, NY, USA. ⁴These authors contributed equally: A. Marguerite, J. Birkbeck, A. Aharon-Steinberg.

*e-mail: david.perello@manchester.ac.uk; eli.zeldov@weizmann.ac.il

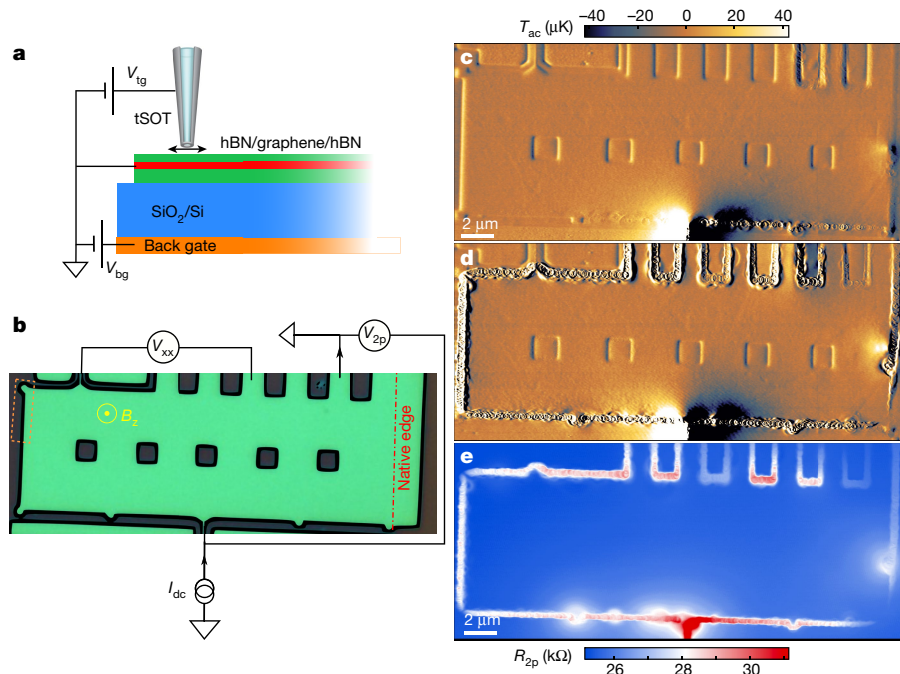


Fig. 1 | Imaging the violation of the quantum Hall topological protection.

a, Schematic of the measurement setup, showing the hBN/graphene/hBN heterostructure and scanning tSOT. **b**, Optical image of device A patterned with several contacts and five square holes in the centre. A d.c. current, I_{dc} , is driven through the narrow bottom constrictions and drained at the top-right contact (arrows) in presence of an applied field $B_z = 1.0$ T at 4.2 K. **c**, Thermal image of $T_{ac}(\mathbf{r})$ in the vicinity of the $\nu = -10$ quantum Hall plateau at $V_{bg} = -6.0$ V ($\nu = -10.7$), $V_{tg} = 8$ V and $I_{dc} = 1.5$ μ A ($R_{2p}I_{dc}^2 = 10$ nW), revealing phonon

emission from individual atomic defects in the form of rings along the bottom-right graphene boundary (see Supplementary Video 1). **d**, Same as **c** but in the quantum Hall plateau transition region at $V_{bg} = -2.5$ V ($\nu = -1.46$) and $I_{dc} = 0.87$ μ A ($R_{2p}I_{dc}^2 = 10$ nW), showing enhanced dissipation along all edges with no visible dissipation in the bulk. **e**, Scanning gate image of $R_{2p}(\mathbf{r})$ acquired simultaneously with **d**, revealing considerable tip-induced enhancement of the two-probe sample resistance along the edges.

corresponding quantized conductance is robust against local perturbations and is determined by the bulk Chern number and the bulk-edge correspondence.

Global transport measurements of our devices show common quantum Hall characteristics—including conductance quantization (see Methods and Extended Data Fig. 4)—which are qualitatively consistent with the above principles. However, when inspected microscopically, we find these principles to be largely violated. A d.c. current $I_{dc} \approx 1$ μ A was injected through a 300-nm-wide constriction at the bottom edge of sample A (Fig. 1b) and was collected at a top-right contact. Figure 1c, d shows the resulting $T_{ac}(\mathbf{r})$ images at two values of the back gate voltage V_{bg} (see Supplementary Information section 1 and Supplementary Video 1 for the range of V_{bg}). There are two features that are independent of V_{bg} : a large thermal gradient near the constriction, which arises from heat that is generated within the constriction and then diffuses through the substrate; and an artificial background signal that outlines the sample topography (see Methods). Notably, V_{bg} -dependent ring-like structures appear along the graphene boundaries, revealing dissipation through phonon emission from individual atomic defects¹¹. When V_{bg} is tuned into the $\nu = -10$ quantum Hall plateau (Fig. 1c, at filling factor $\nu = -10.7$) the dissipation occurs along the bottom edge of the sample, in violation of the first principle listed above. Tuning V_{bg} into the quantum Hall plateau transition region ($\nu = -1.46$, Fig. 1d), dissipation is observed primarily along the edges rather than in the bulk (as demonstrated in particular by the absence of thermal rings at the atomic defects along the inner edges of the five square holes), in violation of the second principle. Moreover, at high filling factors, the dissipation occurs predominantly along the downstream chiral flow direction from the constriction (anticlockwise for holes in Fig. 1c) with a characteristic decay length of about 15 μ m, both in the quantum Hall plateaus and in the plateau transition regions (Supplementary Video 1).

At lower filling factors (Fig. 1d) the dissipation is greatly enhanced in both downstream and upstream directions with no visible chirality, and extends over the entire length of the edges with no apparent decay (Supplementary Information section 1).

Finally, an example of violation of the third principle is demonstrated in Fig. 1e. Topologically protected states should be robust against local perturbations, and hence positive V_{tg} —which depletes holes on a scale much smaller than the sample size—should not affect global transport properties. However, contrary to this, the two-probe resistance R_{2p} of a 30- μ m sample is profoundly affected by a perturbation on a scale of about 50 nm (the tip size). The large increase in $R_{2p}(\mathbf{r})$ occurs only along the graphene boundaries and is observed over a wide range of V_{bg} both at quantum Hall plateaus (Supplementary Video 2) and at plateau transitions (Fig. 1e). It is also of note that the $R_{2p}(\mathbf{r})$ signal is visible along the entire length of the boundaries.

For a closer inspection, we focus on the dashed rectangle in Fig. 1b with a square-shaped protrusion in the top-left corner. The higher-resolution $T_{dc}(\mathbf{r})$ image (Fig. 2a) reveals a disordered heat signal concentrated along two separate contours. The outer contour consists of a series of thermal rings centred along the physical edge of graphene (dashed line). The inner contour, with arc-shaped features, is visible further inside the sample. Critically, the simultaneously acquired scanning gate $R_{2p}(\mathbf{r})$ signal (Fig. 2b) mimics precisely the $T_{dc}(\mathbf{r})$ signal along the inner contour, while showing no response along the outer contour or elsewhere. This difference indicates that the inner and outer contours arise from fundamentally different mechanisms.

To decipher the different mechanisms, we consider a diffusive system in steady state with strong electron–phonon coupling. In this system, dissipation is described by local Joule heating $P(\mathbf{r}) = \mathbf{J}(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r}) = \mathbf{Q}(\mathbf{r})$, where power P is the rate of work W per unit volume, performed by current density \mathbf{J} driven by an electric

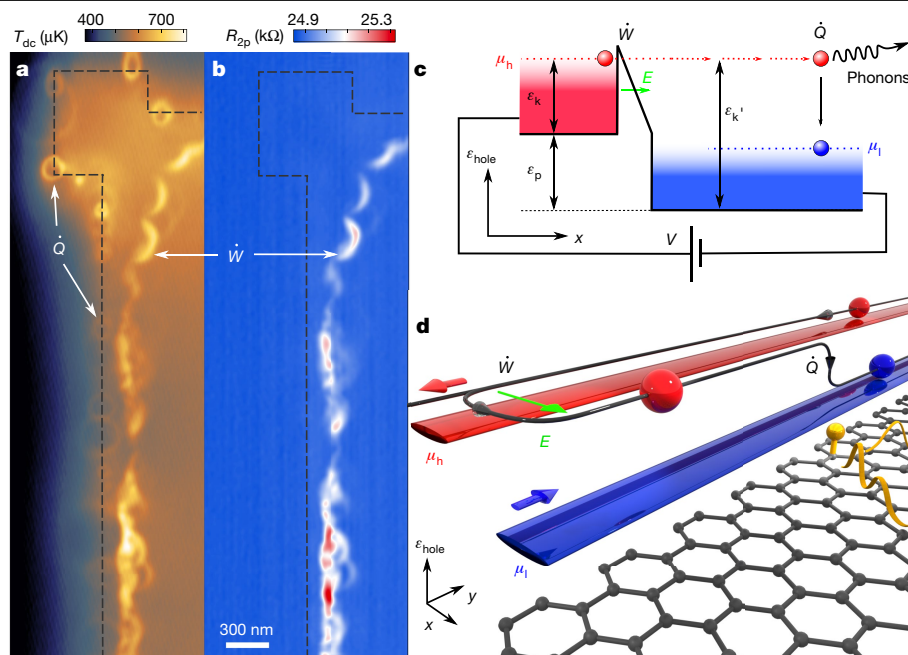


Fig. 2 | Work-generating and entropy-generating processes. **a**, Thermal image of $T_{dc}(\mathbf{r})$ in the corner of sample A (the dashed rectangle in Fig. 1b) at $V_{bg} = -2.1$ V ($\nu \approx -1.9$), $V_{tg} = 1.6$ V and $I_{dc} = 0.62$ μ A, revealing the heat-generating (\dot{Q}) process of phonon emission in form of thermal rings along the graphene boundaries (dashed) and the work-generating (\dot{W}) process of elastic tunnelling between quantum Hall channels along an inner contour. **b**, Simultaneously acquired $R_{2p}(\mathbf{r})$ image showing a \dot{W} process along the inner contour mimicking the corresponding contour in **a**, while displaying no signal associated with the \dot{Q} process along the edges. **c**, Schematic diagram of hole-carrier elastic tunnelling across a potential barrier in a ballistic regime. The work generation \dot{W} occurs only within the barrier at which the electric field (E) (green)

accelerates the carriers transforming the potential energy ϵ_p into excess kinetic energy ϵ_k . The heat \dot{Q} and entropy \dot{Q}/T are generated nonlocally at a remote location through the inelastic scattering of phonons. The difference between the high (μ_h) and low (μ_l) electrochemical potentials is provided by the battery, $\mu_h - \mu_l = eV$. **d**, A schematic cartoon of work generation (\dot{W}) by elastic tunnelling of hole-carriers (spheres) between counterpropagating quantum Hall channels. The transverse electric field (E) (green) performing the work is caused by the difference in the electrochemical potentials μ_h (red) and μ_l (blue). The heat generation (\dot{Q}) occurs at a remote location as a result of phonon emission (light brown) by inelastic carrier scattering off an atomic defect (yellow) at graphene boundaries.

field E . In this case, the work W is transformed into heat Q locally, and hence $\dot{W}(\mathbf{r}) = \dot{Q}(\mathbf{r})$. Conversely, dissipation in a ballistic system can be highly nonlocal, resulting in $\dot{W}(\mathbf{r}) \neq \dot{Q}(\mathbf{r})$, as illustrated in Fig. 2c for elastic tunnelling through a potential barrier. The work generation $\dot{W}(\mathbf{r})$ occurs only where the carriers are accelerated by E within the barrier. Meanwhile, processes that generate heat, $\dot{Q}(\mathbf{r})$, or entropy, $\dot{Q}(\mathbf{r})/T$, occur nonlocally as carriers lose their excess kinetic energy remotely via inelastic scattering of phonons far from the initial barrier.

In general, one should consider three main stages: work generation, equilibration through electron–electron scattering, and heat transfer to the environment through phonon emission. The equilibration process due to electron–electron scattering in the quantum Hall channels has been extensively studied by spectroscopic transport measurements^{12–14}. Such electron–electron scattering results in energy redistribution within the electronic bath, which is undetectable by our technique because no energy is transferred to the phonon bath. In the following, we focus on the first stage of \dot{W} generation and the last stage of \dot{Q} -release into the phonon bath under steady-state conditions, in which the details of the intermediate electron–electron scattering process have no substantial effect. In other words, we address the question of where and how the work is generated and where and how the heat is transferred to the environment.

In an ideal current-carrying ballistic channel, no work-generating processes can take place because there is no potential drop along the channel, $E_{||}(\mathbf{r}) = 0$, and hence $\dot{W}(\mathbf{r}) = 0$. Paradoxically, however, heat can still be generated by the entropy-generating processes, $\dot{Q}(\mathbf{r})$. This is the situation at higher ν , in which—analogue to the tunnel barrier in Fig. 2c—work $\dot{W}(\mathbf{r})$ is performed at the bottom constriction in Fig. 1c

by injecting energetic charge carriers into the quantum Hall edge channels. These chiral carriers flow downstream ballistically and cause nonlocal heating by losing their excess energy to phonons. At low temperatures and in the absence of disorder, electron–phonon coupling is very weak, and as such, phonon emission occurs predominantly through resonant inelastic scattering off single atomic defects along the graphene edges¹¹. These defects form quasi-bound states with sharp energy levels that mediate electron–phonon coupling when in resonance with the incoming charge carriers^{15,16}, giving rise to the \dot{Q} rings observed in Figs. 1c, 2a. Because only forward carrier scattering is allowed in chiral quantum Hall channels, phonon scattering does not affect conductivity and is thus invisible in the $R_{2p}(\mathbf{r})$ image in Fig. 2b and can coexist with full conductance quantization.

At lower fillings, however, markedly different behaviour is observed (Supplementary Information section 1 and Supplementary Video 1). The \dot{Q} rings along the graphene boundaries in Figs. 1d, 2a still reflect nonlocal dissipation, but they are apparently not ‘powered’ by the work generated at the constriction, as evidenced by a lack of observable chirality and of signal decay. Instead, the \dot{W} process occurs along the inner contour in Fig. 2b, where carriers tunnel elastically between neighbouring quantum Hall channels with different electrochemical potential μ , as illustrated schematically in Fig. 2d. The tip positioned at \mathbf{r} modifies the local separation between the channels by its potential V_{tg} . When channels are brought closer together, the tunnelling rate increases and the corresponding backscattering current increases by $\delta I_{bs}(\mathbf{r})$, which in turn increases $R_{2p}(\mathbf{r})$ and generates excess local work at a rate of $\delta \dot{W}(\mathbf{r}) = \delta I_{bs}(\mathbf{r})(\mu_h(\mathbf{r}) - \mu_l(\mathbf{r}))/e$. However, this $\dot{W}(\mathbf{r})$ process is elastic and therefore does not generate local $\dot{Q}(\mathbf{r})$, and no phonons are emitted locally (see Methods and Extended Data Fig. 9). Instead,

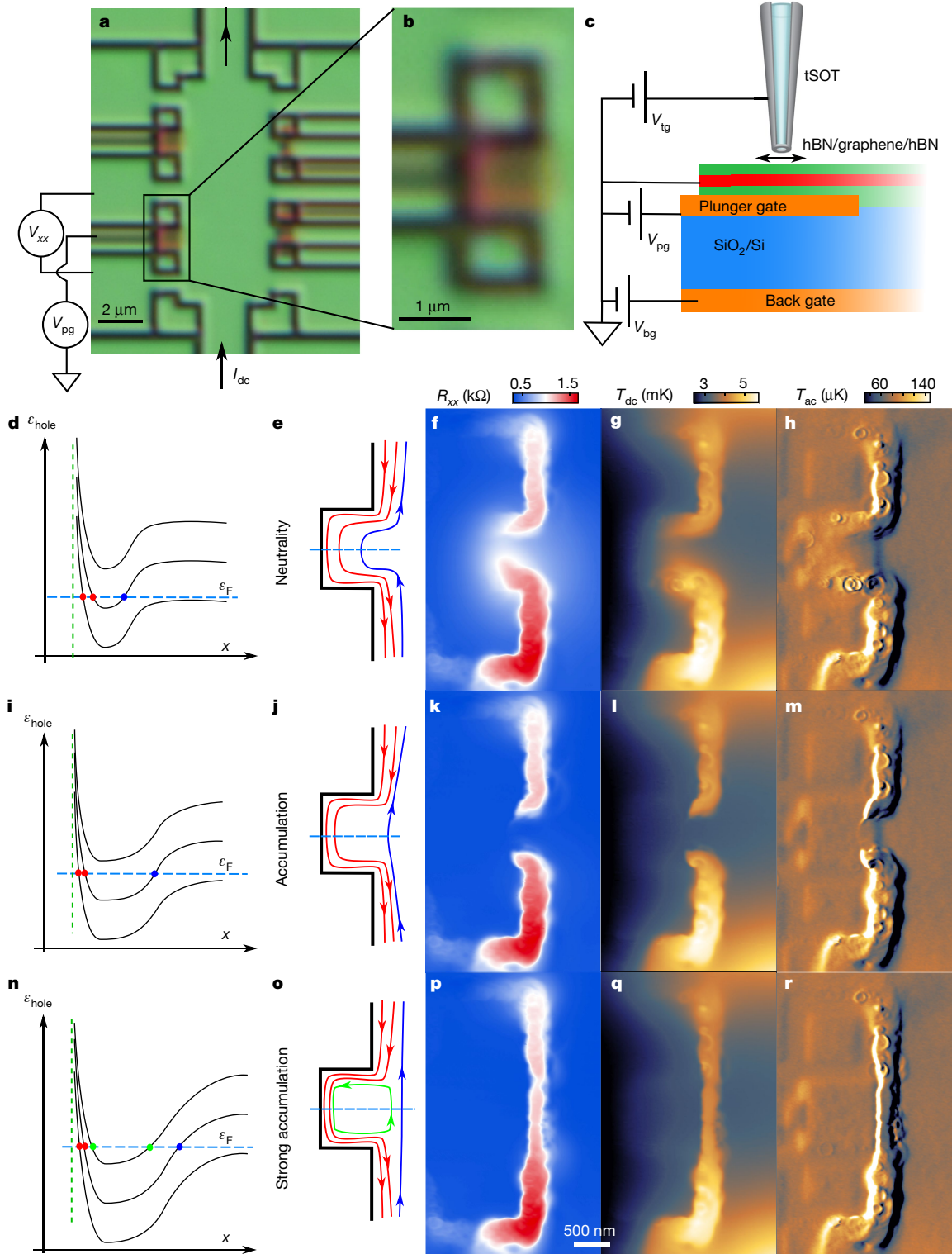


Fig. 3 | Control and separation of the work- and entropy-generating processes. **a**, Optical image of device B with four plunger gates, and a schematic of the measurement circuit (see Supplementary Fig. 2). **b**, Expanded optical image of the plunger-gate region showing the etched trenches (dark) and the underlying plunger gate (orange). **c**, Schematic of the measurement setup, showing the hBN/graphene/hBN heterostructure with a plunger gate and scanning tSOT. **d–h**, Schematic equilibrium Landau-level energy

structure for hole charge carriers (**d**), trajectories of quantum Hall edge channels (**e**), $R_{xx}(\mathbf{r})$ (**f**), $T_{dc}(\mathbf{r})$ (**g**) and $T_{ac}(\mathbf{r})$ (**h**) in the $\nu = -2$ plateau at $B_z = 0.9$ T, $V_{bg} = -1.2$ V ($\nu = -1.73$), $V_{tg} = 3$ V, $V_{pg} = -0.24$ V (neutral plunger gates) and $I_{dc} = 1.75$ μA . **i–m**, Same as **d–h** but for hole-accumulating plunger gates ($V_{pg} = -0.4$ V) that suppress the backscattering $\dot{W}(\mathbf{r})$ process. **n–r**, Same as **d–h** but for strongly hole-accumulating plunger gates ($V_{pg} = -2.0$ V), which creates a $\dot{W}(\mathbf{r})$ contour along the bulk side of the plunger gates.

the backscattered carriers release their excess energy ($\mu_h(\mathbf{r}) - \mu_l(\mathbf{r})$) to phonons elsewhere, predominantly at atomic defects on graphene boundaries. The emitted phonons propagate ballistically, which

increases the overall sample temperature—including at the instantaneous position of the tip, **r**. Because the resulting overall increase in δT and the increase in $R_{2p}(\mathbf{r})$ are both proportional to the tip-induced $\delta \dot{W}(\mathbf{r})$,

the $T_{dc}(\mathbf{r})$ signal in Fig. 2a along the inner contour accurately mimics the $R_{2p}(\mathbf{r})$ signal in Fig. 2b, even though no local $\dot{Q}(\mathbf{r})$ is generated (see Methods and Extended Data Fig. 9).

The above picture, however, raises yet another question. In conventional integer quantum Hall regime, backscattering is prohibited by chirality unless in the presence of counterpropagating channels. Such counterpropagating channels are typically only present at the opposite sample edges, whereas the described $\dot{W}(\mathbf{r})$ process requires proximity between them (Fig. 2d). Our findings, therefore, provide microscopic evidence for the presence of edge reconstruction induced by charge (holes) accumulation along the graphene edges, as has been suggested previously^{7,8,17–20}.

To control this edge reconstruction and investigate its origin, we incorporated plunger gates as described in Fig. 3a–c and Extended Data Figs. 2, 3. As detailed in Fig. 3d, e, band bending due to charge accumulation along the edges creates pairs of counterpropagating quantum Hall channels, which are not topologically protected and exist in addition to the standard topological channels dictated by the bulk-edge correspondence^{7,8,19–21}. These ‘nontopological’ channels provide the means for $\dot{W}(\mathbf{r})$ backscattering and work generation along the entire edge of the sample, as evidenced in Fig. 1d, e, 2b, 3f–h. The presence of band-bending-induced nontopological channels is largely insensitive to the bulk v , and therefore the backscattering $\dot{W}(\mathbf{r})$ occurs for both compressible and incompressible bulk (Supplementary Information section 2, Supplementary Video 2). Moreover, as the nontopological pairs are present at both edges of the sample, the $\dot{W}(\mathbf{r})$ and the resulting nonlocal $\dot{Q}(\mathbf{r})$ show no chiral directionality (Fig. 1d, e). The backscattering rate is determined by the separation between the counterpropagating channels, which we can tune by the plunger-gate potential V_{pg} . Notably, by increasing the hole accumulation the separation between the channels is increased (Fig. 3i, j), which leads to elimination of $\dot{W}(\mathbf{r})$ (Fig. 3k) and of the associated nonlocal heating (Fig. 3l, m) in the region of the plunger gate. Upon further accumulation of holes (see Supplementary Information sections 3, 6 and Supplementary Video 3 for the full sequence) the $\dot{W}(\mathbf{r})$ unexpectedly reappears, but at the bulk side of the plunger gate (Fig. 3n–r), where two copropagating channels (blue and green in Fig. 3o) are formed and hence no backscattering is naively expected. The green channel, however, creates a closed loop and therefore serves as a backscattering mediator between the downstream μ_n (red) and upstream μ_l (blue) channels. Because the green and red channels copropagate along a longer path and are in close proximity due to the steep edge potential (Fig. 3n), the electrochemical potential of the green channel will be close to μ_n . As a result, the overall backscattering rate will be determined by the tunnelling rate between the green and blue channels, explaining the dominant $\dot{W}(\mathbf{r})$ signal along this segment. Note that the patterns along this segment (Fig. 3p–r and Supplementary Fig. 2h, i) are smoother, emphasizing the dominant role of edge disorder in the formation of the complex $\dot{W}(\mathbf{r})$ arc-like patterns along the graphene edges (Figs. 1d, 2a, b). Also, because there are almost no atomic defects in the bulk of graphene¹¹, no $\dot{Q}(\mathbf{r})$ rings are observed along this segment (Supplementary Video 3). The $\dot{Q}(\mathbf{r})$ rings are resolved only along the graphene boundaries (Supplementary Fig. 2), powered by the remote $\dot{W}(\mathbf{r})$, which is consistent with the observed separation of $\dot{Q}(\mathbf{r})$ and $\dot{W}(\mathbf{r})$ contours in Fig. 2.

Even though the plunger gate affects a small region, it considerably influences the global transport (see Methods). A positive (hole-depleting) V_{pg} cuts off the nontopological pairs, increasing R_{xx} (Extended Data Fig. 5d) and forcing the current to bypass the plunger-gate region through the bulk (see Methods and Extended Data Fig. 6). A large tip potential V_{tg} can also cut off the nontopological pairs, as described in Methods, Extended Data Fig. 8, Supplementary Information section 5 and Supplementary Video 5. Note also that the measured $R_{xx}(\mathbf{r})$ is essentially independent of the current (Supplementary Information section 4 and Supplementary Video 4), which rules out possible current-induced quantum Hall breakdown⁴.

The edge reconstruction explains the previously reported discrepancies in the quantum Hall state of graphene⁸ and in other two-dimensional electron gas systems^{21,22}. Although several mechanisms have been proposed^{23,24}, edge accumulation has been mainly ascribed to electrostatic gating^{7,8}, which should lead to symmetric edge accumulation of holes and electrons for p and n doping respectively. We find that at charge neutrality and for both dopings, the accumulation remains hole-type (see Methods and Extended Data Fig. 7), which indicates that the accumulation is predominantly governed by negatively charged impurities. We observe hole-accumulation for both etched and native edges (dashed line in Fig. 1b), despite no chemical exposure for the latter, suggesting that broken bonds at graphene edges become naturally negatively charged. Similar edge accumulation was recently reported in an InAs two-dimensional electron gas²¹. Notably, in proximity-induced superconductivity in graphene and InAs two-dimensional electron gases, the supercurrent was observed to flow preferentially along the edges^{25–28}. Our results may shed light on the underlying mechanism, which is in turn important for studies of topological superconductivity and Majorana physics^{29–31}. Note that the upstream edge channels can undermine the apparent topological protection only if the channels are not well equilibrated⁶. We observe the equilibration length of the upstream channels to be in excess of our sample size of 30 μm (Supplementary Information section 3), which provides a possible explanation for the difficulty in achieving precise quantum Hall quantization in exfoliated graphene devices. Our findings suggest that the detrimental edge reconstruction can be mitigated by passivation or edge-potential engineering^{18,32}. The developed concept of simultaneous work and dissipation imaging, combined with their nanoscale control and spectroscopic analysis, provides a tool for the investigation of microscopic mechanisms of energy loss and scattering in various quantum and topological systems and in operational electronic nanodevices.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1704-3>.

- Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
- Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057–1110 (2011).
- Halbatal, D. et al. Nanoscale thermal imaging of dissipation in quantum systems. *Nature* **539**, 407–410 (2016).
- Tzalenchuk, A. et al. Towards a quantum resistance standard based on epitaxial graphene. *Nat. Nanotechnol.* **5**, 186–189 (2010).
- Weis, J. & von Klitzing, K. Metrology and microscopic picture of the integer quantum Hall effect. *Philos. Trans. R. Soc. A* **369**, 3954–3974 (2011).
- Büttiker, M. Absence of backscattering in the quantum Hall effect in multiprobe conductors. *Phys. Rev. B* **38**, 9375–9389 (1988).
- Silvestrov, P. G. & Efetov, K. B. Charge accumulation at the boundaries of a graphene strip induced by a gate voltage: Electrostatic approach. *Phys. Rev. B* **77**, 155436 (2008).
- Cui, Y.-T. et al. Unconventional correlation between quantum Hall transport quantization and bulk state filling in gated graphene devices. *Phys. Rev. Lett.* **117**, 186601 (2016).
- Vasyukov, D. et al. A scanning superconducting quantum interference device with single electron spin sensitivity. *Nat. Nanotechnol.* **8**, 639–644 (2013).
- Eriksson, M. A. et al. Cryogenic scanning probe characterization of semiconductor nanostructures. *Appl. Phys. Lett.* **69**, 671–673 (1996).
- Halbatal, D. et al. Imaging resonant dissipation from individual atomic defects in graphene. *Science* **358**, 1303–1306 (2017).
- Altımiras, C. et al. Tuning energy relaxation along quantum Hall channels. *Phys. Rev. Lett.* **105**, 226804 (2010).
- Venkatachalam, V., Hart, S., Pfeiffer, L., West, K. & Yacoby, A. Local thermometry of neutral modes on the quantum Hall edge. *Nat. Phys.* **8**, 676–681 (2012).
- Itoh, K. et al. Signatures of a nonthermal metastable state in copropagating quantum Hall edge channels. *Phys. Rev. Lett.* **120**, 197701 (2018).
- Tikhonov, K. S., Gornyi, I. V., Kachorovskii, V. Y. & Mirlin, A. D. Resonant supercollisions and electron-phonon heat transfer in graphene. *Phys. Rev. B* **97**, 085415 (2018).
- Kong, J. F., Levitov, L., Halbatal, D. & Zeldov, E. Resonant electron-lattice cooling in graphene. *Phys. Rev. B* **97**, 245416 (2018).

17. Chae, J. et al. Enhanced carrier transport along edges of graphene devices. *Nano Lett.* **12**, 1839 (2012).
18. Panchal, V. et al. Visualisation of edge effects in side-gated graphene nanodevices. *Sci. Rep.* **4**, 5881 (2014).
19. Vera-Marun, I. J. et al. Quantum Hall transport as a probe of capacitance profile at graphene edges. *Appl. Phys. Lett.* **102**, 013106 (2013).
20. Barraud, C. et al. Field effect in the quantum Hall regime of a high mobility graphene wire. *J. Appl. Phys.* **116**, 073705 (2014).
21. Akiho, T., Irie, H., Onomitsu, K. & Muraki, K. Counterflowing edge current and its equilibration in quantum Hall devices with sharp edge potential: Roles of incompressible strips and contact configuration. *Phys. Rev. B* **99**, 121303 (2019).
22. Ma, E. Y. et al. Unexpected edge conduction in mercury telluride quantum wells under broken time-reversal symmetry. *Nat. Commun.* **6**, 7252 (2015).
23. Shtanko, O. & Levitov, L. Robustness and universality of surface states in Dirac materials. *Proc. Natl Acad. Sci. USA* **115**, 5908–5913 (2018).
24. Akhmerov, A. R. & Beenakker, C. W. J. Boundary conditions for Dirac fermions on a terminated honeycomb lattice. *Phys. Rev. B* **77**, 085423 (2008).
25. Allen, M. T. et al. Spatially resolved edge currents and guided-wave electronic states in graphene. *Nat. Phys.* **12**, 128–133 (2016).
26. Amet, F. et al. Supercurrent in the quantum Hall regime. *Science* **352**, 966–969 (2016).
27. Zhu, M. J. et al. Edge currents shunt the insulating bulk in gapped graphene. *Nat. Commun.* **8**, 14552 (2017).
28. Indolese, D. I. et al. Signatures of van Hove singularities probed by the supercurrent in a graphene–hBN superlattice. *Phys. Rev. Lett.* **121**, 137701 (2018).
29. Pribiag, V. S. et al. Edge-mode superconductivity in a two-dimensional topological insulator. *Nat. Nanotechnol.* **10**, 593–597 (2015).
30. Lee, G.-H. et al. Inducing superconducting correlation in quantum Hall edge states. *Nat. Phys.* **13**, 693–698 (2017).
31. de Vries, F. K. et al. h/e Superconducting quantum interference through trivial edge states in InAs. *Phys. Rev. Lett.* **120**, 047702 (2018).
32. Ribeiro-Palau, R. et al. High-quality electrostatically defined Hall bars in monolayer graphene. *Nano Lett.* **19**, 2583–2587 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Device fabrication

Monolayer graphene heterostructures were fabricated by exfoliating natural graphite and hBN flakes onto oxidized silicon wafers (290 nm of SiO₂) and stacking via a polymer stamping method³³. This method can achieve contamination-free areas limited only by the size of the hBN. Our devices comprised a relatively thick bottom hBN crystal (>30 nm) with a thinner (10–20 nm) crystal covering the monolayer graphene. We intentionally misaligned the graphene edges with respect to either hBN (>5°) to avoid any superlattice effects. For samples B and C incorporating a plunger gate, the bottom gate structures were first patterned and metallized with Cr/Au (1 nm/9 nm), followed by transferring of the annealed hBN/graphene/hBN heterostructure.

We used electron beam lithography (Raith EBPG5200) with a bilayer (A3 495K, A3 950K) polymethyl methacrylate (PMMA) mask to define both the contact location and sample geometry. To improve either contact resistance or edge sharpness, we incorporated two different CHF₃ and O₂ reactive ion etching (RIE) recipes for each of the contacts and mesa definition. Contacts were defined by mixed chemical/physical etching (5 W RIE, 150 W inductively-coupled plasma (ICP)) to improve selectivity for hBN over PMMA, and thus allowing etching and metallization in a single step. Mesa etching incorporated a physical RIE process (20 W RIE, 0 W ICP), followed by a weak Ar/O₂ RIE etch to remove the residual exposed graphene step at the edges²⁷. Contact metallization was achieved via e-beam evaporation of Cr/Au (1 nm/70 nm) and standard liftoff procedure. Finally, before scanning, the samples were soaked in a tetramethylammonium hydroxide (TMAH)-based alkaline developer (MIF-319) to remove residual PMMA resist from the surface of the heterostructure. These fabrication procedures are known to produce samples with high electron mobility and ballistic transport, with a momentum-relaxing mean free path limited by the sample dimensions^{11,34}.

Sample A had a main chamber of 30 × 10 μm² (Fig. 1b, Extended Data Fig. 1). Constrictions of 300 and 200 nm width at the bottom and top-left edges were designed to allow injection of energetic carriers into the quantum Hall edge channels. We also etched a series of 1.5 × 1.5 μm² holes in the centre of the main chamber in order to visualize dissipation in the centre of the device by detecting \dot{Q} rings along the inner edges of the holes as demonstrated previously³. Protruding 500 × 250 nm² rectangles on the lower and left edges of the mesa served as poking pads to facilitate tSOT scanning height control using tuning fork feedback. For the native edge (dashed line in Extended Data Fig. 1), we chose a straight region of the flake, which indicates one of the main crystallographic cleaving directions of graphene. Samples B (Fig. 3a, Extended Data Figs. 2, 5a) and C (Extended Data Fig. 3) were defined by 200-nm-wide trenches, with the main chamber size of 10 × 4 μm² and 13 × 5 μm² respectively. The typical two-probe resistances in zero magnetic field at 4.2 K in sample A were 1–2 kΩ and approximately 8 kΩ through the constriction, and 7–8 kΩ in samples B and C.

The hBN-encapsulated graphene samples residing on Si/SiO₂ substrates were glued to Au-plated G10 chip carrier using silver paint. Gold wires for electrical contacts were glued to lithographically defined bonding pads using silver epoxy cured at room temperature for 4 h. The scanning probe microscope with the sample resides in a brass vacuum chamber that is immersed in liquid He at 4.2 K. The chamber is filled with He exchange gas at around 60 mbar pressure providing thermal coupling between the tSOT and the sample³ and a good thermal contact between the entire microscope and the liquid He bath.

In the \dot{Q} process, phonons are emitted from atomic defects and propagate ballistically throughout the sample. The measured local temperature increase reflects the local energy density of the excess phonons. In the microscopic vicinity of the phonon emitter the excess temperature is determined by the emission rate and the distance from the emitter^{3,11}, and is essentially independent of the thermal resistance

between the sample and the environment. The latter determines the overall average temperature increase of the entire sample dictated by the overall power dissipated in the sample, which will appear as a small global increase in T_{dc} .

tSOT fabrication and measurement schemes

The Pb tSOTs were fabricated as previously described⁹ with effective diameters of 43, 57 and 50 nm for samples A, B, and C respectively. All the presented scanning studies were carried out at 4.2 K in presence of He exchange gas at a pressure of around 60 mbar in magnetic fields of $B_z = 1$ T for sample A, 0.9 T for sample B and 1.2 T for sample C. At these conditions the tSOTs displayed thermal noise of down to 0.75 μK/Hz^{1/2}. For height control the tSOT was attached to a quartz tuning fork as described in refs.^{35,36} and was electrically excited at a resonance frequency of around 35 kHz. The scanning was performed at a constant height of 25–50 nm above the surface of the top hBN. The tuning fork was vibrated along the \hat{x} direction, causing the tSOT to vibrate with it with a controllable amplitude x_{ac} , with root mean square amplitude values ranging between 5–30 nm. The typical scanning parameters were: pixel size 6–30 nm, acquisition time 10–60 ms per pixel, with image sizes from 100 × 100 to 500 × 500 pixels per image. The imaging was performed using four modalities:

d.c. thermal imaging $T_{dc}(\mathbf{r})$. A current I_{dc} is applied to the sample chopped by a square wave at a frequency of 94 Hz and the corresponding thermal map $T_{dc}(\mathbf{r})$ of the sample is acquired using lock-in amplifier locked to the chopping frequency. As a result, the $T_{dc}(\mathbf{r})$ image provides a map of the current-induced local temperature increase in the sample.

Second harmonic thermal imaging $T_{2f}(\mathbf{r})$. Similar to the $T_{dc}(\mathbf{r})$, instead of the d.c. current, a sinusoidal a.c. current I_{ac} is applied to the sample at frequency f and the corresponding thermal map $T_{2f}(\mathbf{r})$ of the sample is acquired by a lock-in amplifier locked to the second harmonic frequency $2f$.

a.c. thermal imaging $T_{ac}(\mathbf{r})$ at tuning fork frequency. Using the fact that the tSOT is mounted on the tuning fork, we also measure $T_{ac}(\mathbf{r}) \approx x_{ac} \partial T_{dc}(\mathbf{r}) / \partial x$ using a lock-in amplifier locked onto the excitation frequency of the tuning fork. The advantage of this mode is that it enhances the visibility of the sharp local features relative to the smooth background, and in particular of the $\dot{Q}(\mathbf{r})$ rings. The $T_{ac}(\mathbf{r})$ signal, however, contains another small component due to the few-mK self-heating of the tSOT induced by the measurement current applied to it. In the absence of a current in the sample, the self-temperature of the tSOT is slightly position dependent due to the cooling of the tip by the sample (see supplementary figure Se in ref.³ and its accompanying description) which results in a small contrast between the hBN regions in the sample and the etched trenches. The $T_{ac}(\mathbf{r})$ contains a contribution from the gradient of this contrast, visible as contour lines of the etched hBN mesas in Fig. 1c, d. These contours, which are visible also in the absence of a current in the sample, are highly beneficial for navigation and visualization of the sample with high precision. Because this signal is current-independent, it is absent in the $T_{dc}(\mathbf{r})$ images which measure only the temperature increase due to the chopped current I_{dc} .

Scanning gate imaging $R(\mathbf{r})$. By applying a voltage V_{tg} between the tSOT and the sample we carry out scanning gate imaging using a method similar to those reported previously^{10,37–44} simultaneously with the thermal imaging. In particular, the voltage difference V between a pair of sample contacts is measured using a lock-in amplifier locked to the chopping frequency of the current I_{dc} , and then $R(\mathbf{r}) = V(\mathbf{r})/I_{dc}$ is plotted against the tip location \mathbf{r} . In this manner either two-probe, $R_{2p}(\mathbf{r})$, or four-probe, $R_{xx}(\mathbf{r})$, tip-position dependent resistance values are attained.

Transport characteristics

Four-point transport characterization measurements were performed using standard lock-in techniques at 5.4 Hz. Extended Data Fig. 4 shows the Landau fans of samples B and C. From the slopes of the resistance minima we extract the capacitance $C = \frac{ev}{\phi_0} \frac{\partial B_z}{\partial V_{bg}}|_{v \in \{\pm 2, \pm 6, \pm 10, \dots\}} = 1.005 \times 10^{-8} \text{ F cm}^{-2} = 6.27 \times 10^{10} e \text{ per cm}^2 \text{ V}$ for sample B and $0.85 \times 10^{-8} \text{ F cm}^{-2}$ for sample C, where e is the elementary charge. Using the zero field resistivity data ρ_{xx} we derive the mobility $\mu = 1/(n_e e \rho_{xx}) = 7.0 \times 10^5 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and the mean free path $l_{\text{mfp}} = \frac{1}{2k_F \rho_{xx}} \frac{h}{e^2} = 4.5 \text{ }\mu\text{m}$ at $n_e = 2.8 \times 10^{11} \text{ cm}^{-2}$ for sample B, and $\mu = 2.13 \times 10^5 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $l_{\text{mfp}} = 1.19 \text{ }\mu\text{m}$ at $n_e = 2.4 \times 10^{11} \text{ cm}^{-2}$ for sample C; h is Planck's constant. Here $n_e = C(V_{bg} - V_{bg}^{\text{CNP}})/e$ is the carrier density, $v = n_e \phi_0 / B_z$ is the filling factor, $V_{bg}^{\text{CNP}} = -0.6 \text{ V}$ is the charge neutrality point (-1.85 V for sample C), $k_F = \sqrt{\pi n_e}$ is the graphene Fermi wavevector, and $\phi_0 = h/e$. Because of its unconventional geometry and limited working contacts, we could not properly measure the Landau fan diagram of sample A to extract its mobility and mean free path. However, R_{xx} measurements show a similar behaviour, from which we extract V_{bg}^{CNP} and the approximate filling factors. Note that Figs. 1 and 2 were acquired at different cool downs, resulting in a shift in V_{bg}^{CNP} .

Extended Data Fig. 5 describes the effect of plunger gate on the global transport. Because the size of the plunger gate is much smaller than the sample, it should naively have no measurable effect in a topologically protected state. Extended Data Fig. 5b shows that, when analysed on a linear scale, the variations in σ_{xx} and R_{xx} with V_{pg} may not appear to be very substantial; however, on a logarithmic scale (Extended Data Fig. 5c) variations of up to two orders of magnitude in R_{xx} are visible, in particular around $v = \pm 2$ plateaus. Extended Data Fig. 5d shows R_{xx} at $V_{bg} = -1 \text{ V}$ in the vicinity of the $v = -2$ plateau plotted against V_{pg} , displaying a stepwise increase from $R_{xx} \approx 100 \text{ }\Omega$ to over $4 \text{ k}\Omega$ for $V_{pg} \geq -0.23 \text{ V}$. Hole edge accumulation creates additional pairs of counterpropagating nontopological channels that reduce the sample resistivity. When the hole accumulation is depleted by applying a positive V_{pg} , the highly conductive nontopological channels are cut off (Extended Data Fig. 6a), leading to the observed sharp increase in R_{xx} . Hole edge accumulation is present also for n-doping of graphene and is visible in Extended Data Fig. 5c up to $V_{bg} \approx 2.3 \text{ V}$, above which the V_{pg} dependence decreases considerably, indicating the dominant contribution of negatively charged impurities to the hole edge accumulation (see Methods).

Cutting off the nontopological edge channels by plunger gate V_{pg}

The enhanced conductivity of the nontopological pairs of channels due to hole accumulation provides low resistance paths for the current flow. This accumulation, however, can be locally depleted by the plunger gate with $V_{pg} > -0.23 \text{ V}$, as indicated by transport measurements in Extended Data Fig. 5d. In this case, the nontopological pairs are cut off (Extended Data Fig. 6a), causing a pronounced increase in the global R_{xx} . Notably, in this situation the current that is carried by the nontopological channels is partially forced to flow through the bulk in the cut off segment. A depleting V_{pg} increases the local bulk resistivity under the tip, therefore enhancing $R_{xx}(\mathbf{r})$ (as observed by the diffused red blob in Extended Data Fig. 6b and Supplementary Video 3), revealing the current path through the bulk. Note that the topologically protected channel remaining in the depleted region (red in Extended Data Fig. 6a) still carries current; however, the resulting potential drop that develops across the plunger gate region imposes a parallel partial conduction through the bulk. The current that flows in the topological channel, however, cannot be visualized by $R_{xx}(\mathbf{r})$ because the downstream flowing carriers there cannot backscatter to another channel and do not perform work. However, these carriers can still lose their excess energy by phonon emission at the atomic defects, giving rise to the \dot{Q} rings along the graphene boundaries as observed in Extended Data Fig. 6c.

Because the nontopological channels are cut off, this case provides an insight into work and dissipation that should occur in their absence. Extended Data Fig. 6b shows that the carriers tunnel between the edge states through the bulk as expected in the quantum Hall plateau transition regions. When the nontopological edges are present, however, they shunt the bulk by providing low-resistance paths for carrier backscattering and hence hardly any work and dissipation are observed in the bulk of the sample even in the plateau transition regions.

Hole accumulation at the graphene edges for n-doped bulk

Edge charge accumulation causes the formation of nontopological pairs of channels that provide a low-resistance path for current flow. These nontopological pairs can be cut off by a depleting plunger gate leading to an increase in R_{xx} . In the case of hole accumulation this is demonstrated by applying a positive V_{pg} , while a negative V_{pg} does not affect R_{xx} substantially because increasing local accumulation only lowers the local resistance, which is already relatively low (Extended Data Fig. 5d). Extended Data Fig. 5c indeed shows that for negative V_{bg} (p-doped bulk) a depleting (positive) V_{pg} increases R_{xx} . If the edge accumulation would be solely caused by backgate electrostatics^{7,20,45,46}, the situation would be inverted for n-doped bulk; namely, a negative V_{pg} would deplete the electron accumulation along the edges thus increasing R_{xx} .

However, Extended Data Fig. 5c shows that it is not the case and R_{xx} is increased by a positive V_{pg} even in the n-doped region (for $V_{bg} \lesssim 2.3 \text{ V}$). This implies that for moderate n-doping of the bulk, the edges still remain p-doped, as confirmed microscopically in Extended Data Fig. 7. Here the $R_{xx}(\mathbf{r})$ scans for n-doped bulk show that—similarly to the case of p-doped bulk—a positive, rather than a negative, V_{pg} increases the resistance along the edges. This hole edge accumulation⁴⁷ is clearly resolved in the vicinity of $v = 2$ and 6 plateaus as shown in Extended Data Fig. 7a, b.

Demonstration of the elastic \dot{W} scattering and nonlocal heating

We present here a more detailed evidence that the $\dot{W}(\mathbf{r})$ process is predominantly elastic. For this we first summarize the effect of the tip potential V_{tg} , which is crucial for revealing the described phenomena. At flat band conditions $V_{tg} = V_{tg}^{\text{FB}} \approx 0 \text{ V}$ the tip has no influence on the sample. In this case $R_{xx}(\mathbf{r})$ is fixed independent of the tip position (Extended Data Fig. 8a) and therefore the $\dot{W}(\mathbf{r})$ processes cannot be imaged (Extended Data Fig. 9a). Nonetheless, the $T_{dc}(\mathbf{r})$ signal due to $\dot{Q}(\mathbf{r})$ processes is present, but the phonons—even though they are being emitted predominantly at resonant states at atomic defects—propagate ballistically throughout the sample. As a result, the $T_{dc}(\mathbf{r})$ profiles are smooth (Extended Data Fig. 9b) and hence the atomic-scale $\dot{Q}(\mathbf{r})$ sources cannot be resolved individually.

Upon applying a finite V_{tg} , however, both the $\dot{Q}(\mathbf{r})$ and $\dot{W}(\mathbf{r})$ processes can be clearly identified. The individual $\dot{Q}(\mathbf{r})$ sources are revealed through the formation of the temperature rings around them (for example, Supplementary Fig. 2), which reflect the loci of the tip positions at which the tSOT potential V_{tg} brings the localized resonant electronic states of the defects to the Fermi energy¹¹. Similarly, applying a small positive V_{tg} allows imaging of the locations at which $\dot{W}(\mathbf{r})$ is present and hence revealing the locations of the nontopological channels by shifting them slightly closer to each other (Extended Data Fig. 8b), thus enhancing the local elastic tunnelling rates between them by $\delta \dot{W}(\mathbf{r})$ and increasing the $R_{xx}(\mathbf{r})$. The observed rich patterns of $\delta \dot{W}(\mathbf{r})$ reflect the intricate trajectories of the quantum Hall channels and the variations in the local separation between them due to electrostatic disorder. In particular, it reveals the nontopological channels at the inner edge of the plunger gate in Extended Data Fig. 9c. A further increase of the depleting V_{tg} can entirely cut off the nontopological pairs (Extended Data Fig. 8c) and even induce an n-doped region under the tip (Extended Data Fig. 8d).

Throughout the paper we refer to the \dot{W} process of carrier tunnelling between the channels as a purely elastic process with no local phonon emission, in which case all the \dot{Q} processes are nonlocal (Fig. 2d). One can also consider a higher-order inelastic tunnelling between the channels, in which a phonon is emitted concurrently with tunnelling, resulting in a local \dot{Q} at the tunnelling location. We can discern the two cases by considering the perturbation induced by the tip potential. A weakly perturbing tip has two effects: enhancing backscattering $\delta\dot{W}(\mathbf{r})$, thus revealing the locations of $\dot{W}(\mathbf{r})$ processes through $R_{xx}(\mathbf{r})$ imaging (Extended Data Fig. 9c); and enhancing heating (either local or nonlocal) as a result of the enhanced $\delta\dot{W}(\mathbf{r})$. Figure 2, Supplementary Fig. 2 and Supplementary Video 3 clearly demonstrate that the \dot{Q} rings at atomic defects along the graphene boundaries reflect nonlocal heating. However, the observed enhanced temperature signal along the quantum Hall channels (for example, on the bulk-side edge of the plunger gates in Extended Data Fig. 9d) could reflect either local heating due to higher-order inelastic carrier tunnelling between the quantum Hall channels or a nonlocal heating due to phonons emitted at remote locations causing overall temperature increase detected as an enhanced $T_{dc}(\mathbf{r})$ at the instantaneous position of the tip \mathbf{r} . These two possibilities are hard to distinguish by inspecting only the tip-perturbing images such as those in Extended Data Fig. 9d. A non-perturbing tip, by contrast, performs only one function: imaging the unperturbed temperature distribution. If the tunnelling between quantum Hall channels is elastic then the heating is nonlocal and thus the maximum of $T_{dc}(\mathbf{r})$ should occur along the graphene boundaries where the phonons are emitted at the atomic defects regardless of where $\dot{W}(\mathbf{r})$ occurs. If, on the other hand, the tunnelling is inelastic, a peak in $T_{dc}(\mathbf{r})$ should occur along the $\dot{W}(\mathbf{r})$ contours. Usually the $\dot{W}(\mathbf{r})$ contours are located close to the graphene boundaries, but near the sample corners they can be considerably shifted towards the bulk (Fig. 2a, b) or, alternatively, we can shift them in a controllable manner using the plunger gate (Extended Data Fig. 9).

Extended Data Fig. 9e, f presents three $T_{dc}(\mathbf{r})$ profiles along the colour lines in Extended Data Fig. 9b for the case of a non-perturbing tip. The green profile shows that $T_{dc}(\mathbf{r})$ is maximal along the graphene boundaries with a slowly decaying tail into the bulk of the sample due to ballistic phonon propagation. The red and blue profiles show that the slow tails of $T_{dc}(\mathbf{r})$ originating from the three boundaries overlap, resulting in a plateau-like profile in the sample protrusion region. The key observation, however, is that the blue profile in Extended Data Fig. 9e shows no peak in $T_{dc}(\mathbf{r})$ at the location of the $\dot{W}(\mathbf{r})$ contour on the bulk side of the plunger gate as revealed in Extended Data Fig. 9c, d. These results demonstrate that the $\dot{W}(\mathbf{r})$ scattering is predominantly elastic and that the $\dot{Q}(\mathbf{r})$ dissipation is predominantly nonlocal occurring at the atomic defects at graphene boundaries.

Data availability

Data supporting the findings of this study are available within the article and its Supplementary Information files and from the corresponding authors upon reasonable request.

33. Pizzocchero, F. et al. The hot pick-up technique for batch assembly of van der Waals heterostructures. *Nat. Commun.* **7**, 11894 (2016).
34. Ella, L. et al. Simultaneous voltage and current density imaging of flowing electrons in two dimensions. *Nat. Nanotechnol.* **14**, 480–487 (2019).
35. Finkler, A. et al. Self-aligned nanoscale SQUID on a tip. *Nano Lett.* **10**, 1046–1049 (2010).
36. Finkler, A. et al. Scanning superconducting quantum interference device on a tip for magnetic imaging of nanoscale phenomena. *Rev. Sci. Instrum.* **83**, 073702 (2012).
37. Paradiso, N. et al. Spatially resolved analysis of edge-channel equilibration in quantum Hall circuits. *Phys. Rev. B* **83**, 155305 (2011).
38. Garcia, A. G. F., König, M., Goldhaber-Gordon, D. & Todd, K. Scanning gate microscopy of localized states in wide graphene constrictions. *Phys. Rev. B* **87**, 085446 (2013).
39. Pascher, N. et al. Imaging the conductance of integer and fractional quantum Hall edge states. *Phys. Rev. X* **4**, 011014 (2014).
40. Bhandari, S. et al. Imaging cyclotron orbits of electrons in graphene. *Nano Lett.* **16**, 1690–1694 (2016).
41. Braem, B. A. et al. Investigating energy scales of fractional quantum Hall states using scanning gate microscopy. *Phys. Rev. B* **93**, 115442 (2016).
42. Dou, Z. et al. Imaging bulk and edge transport near the Dirac point in graphene moiré superlattices. *Nano Lett.* **18**, 2530–2537 (2018).
43. Herbschleb, E. D. et al. Direct imaging of coherent quantum transport in graphene p–n–p junctions. *Phys. Rev. B* **92**, 125414 (2015).
44. Schnez, S. et al. Imaging localized states in graphene nanostructures. *Phys. Rev. B* **82**, 165445 (2010).
45. Bischoff, D. et al. Localized charge carriers in graphene nanodevices. *Appl. Phys. Rev.* **2**, 031301 (2015).
46. Vasko, F. T. & Zozoulenko, I. V. Conductivity of a graphene strip: width and gate-voltage dependencies. *Appl. Phys. Lett.* **97**, 092115 (2010).
47. Woessner, A. et al. Near-field photocurrent nanoscopy on bare and encapsulated graphene. *Nat. Commun.* **7**, 10783 (2016).

Acknowledgements We thank G. Zhang, I. V. Gornyi, A. D. Mirlin and Y. Gefen for discussions and theoretical analysis, M. E. Huber for SOT readout setup, and M. L. Rappaport for technical assistance. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant number 785971), by the Israel Science Foundation (ISF; grant number 921/18), by the Minerva Foundation with funding from the Federal German Ministry of Education and Research, by the German-Israeli Foundation (GIF), by the Weizmann–UK Making Connections Program, and by Manchester Graphene-NOWNANO CDT EP/L-1548X. E.Z. acknowledges the support of the Leona M. and Harry B. Helmsley Charitable Trust grant 2018PG-ISL006.

Author contributions A.M., A.A.-S., D.H., I.M. and E.Z. conceived the experiments. J.B. and D.J.P. conceived and fabricated the samples. A.M. and A.A.-S. carried out the measurements and data analysis. D.H. and I.M. performed preliminary studies. K.B. and Y.M. fabricated the SOTs and the tuning fork feedback. A.M., E.Z., A.A.-S., D.J.P., J.B. and A.K.G. wrote the manuscript. All authors participated in discussions and writing of the manuscript.

Competing interests The authors declare no competing interests.

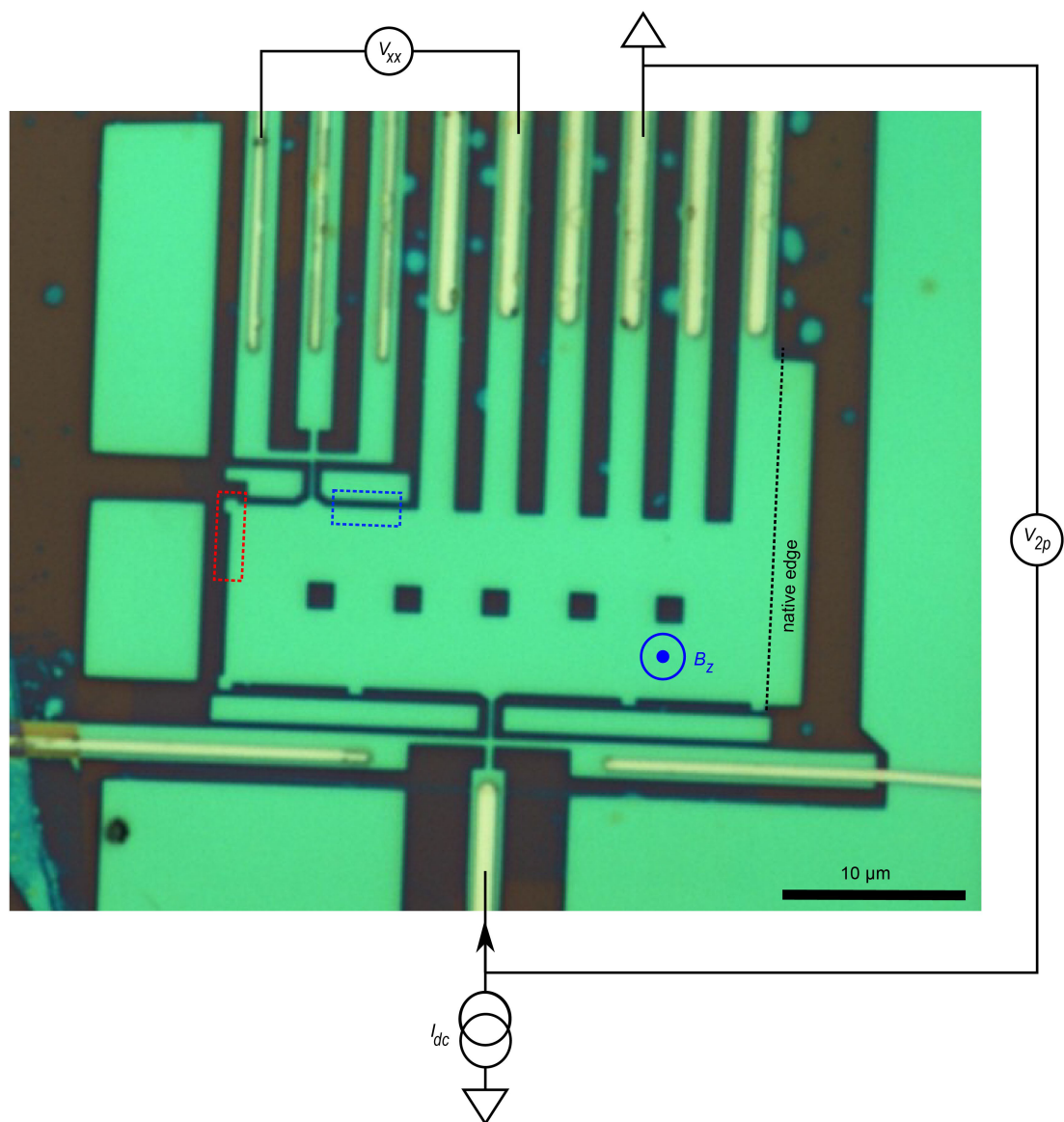
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1704-3>.

Correspondence and requests for materials should be addressed to D.J.P. or E.Z.

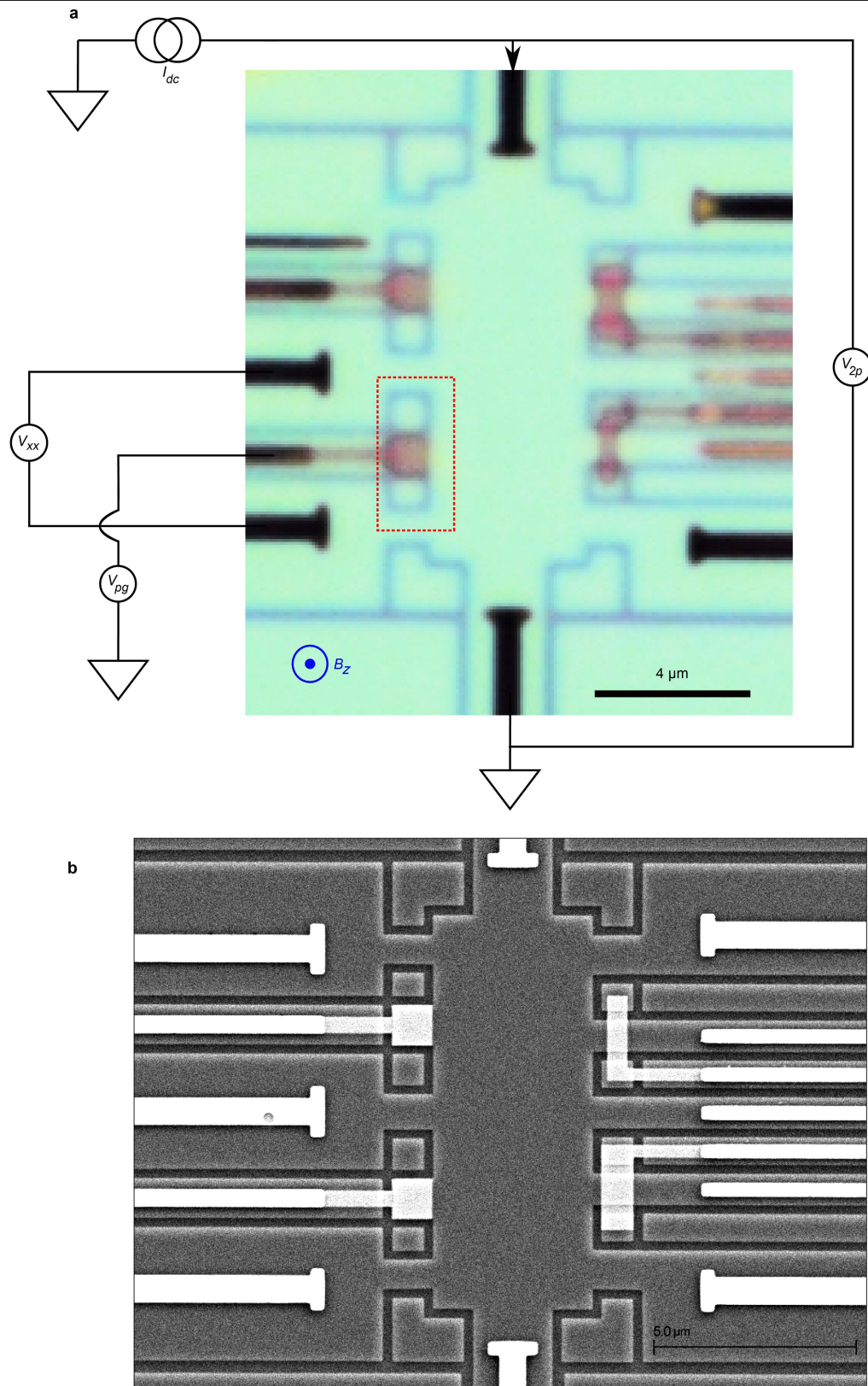
Peer review information *Nature* thanks Xi Lin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



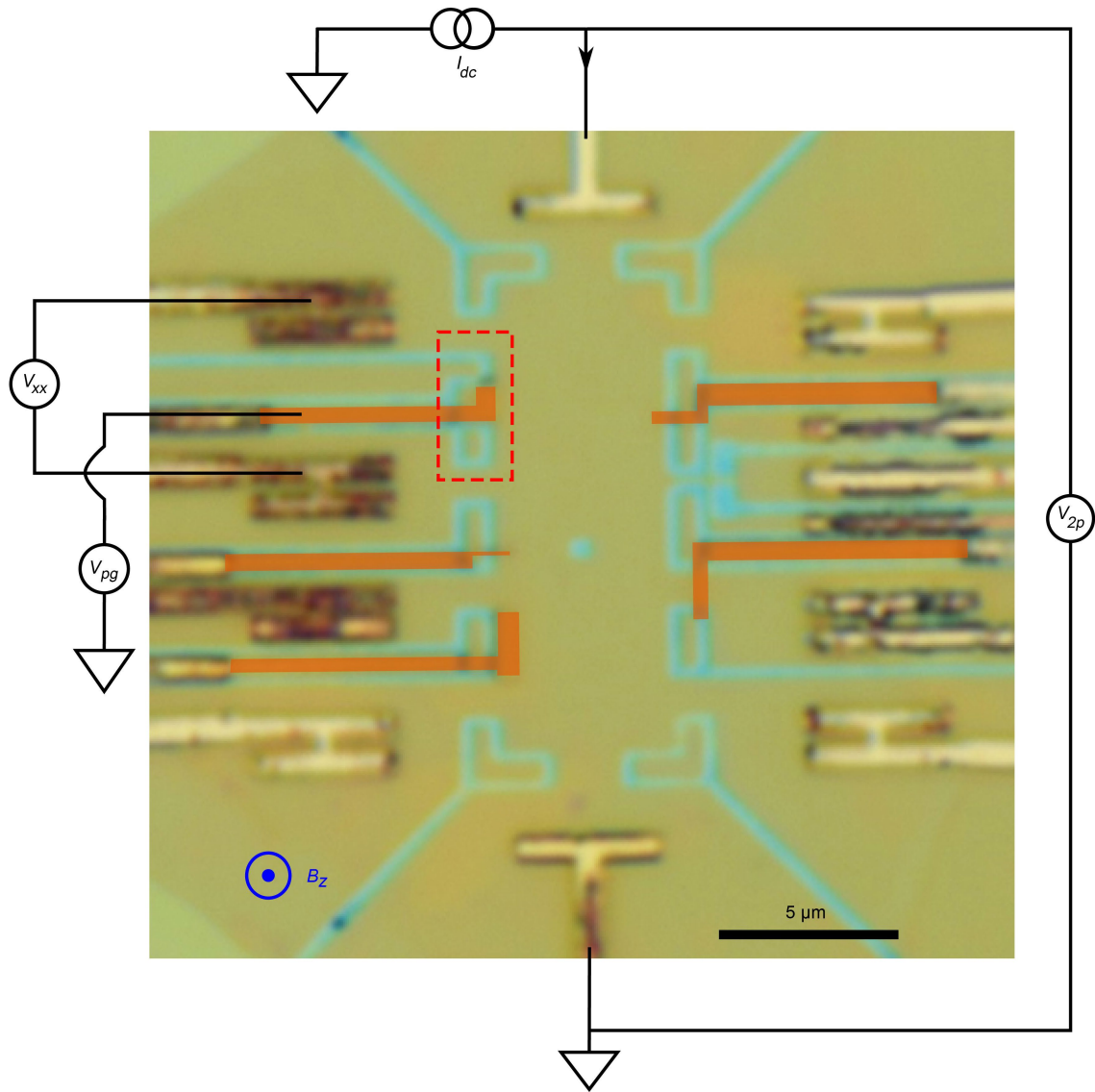
Extended Data Fig. 1 | Optical image of sample A. Shown are the hBN/graphene/hBN heterostructure (green), the etched regions exposing the SiO₂/Si substrate (dark) and the metal contacts (yellow). The dashed rectangles mark the regions shown in Fig. 2 (red) and Supplementary Video 2 (blue). The

current is applied to the bottom constriction and drained at the top contact, and the corresponding voltages V_{xx} and V_{2p} are measured in the scanning-gate mode. The dashed line on the right shows the native edge of graphene encapsulated in hBN.



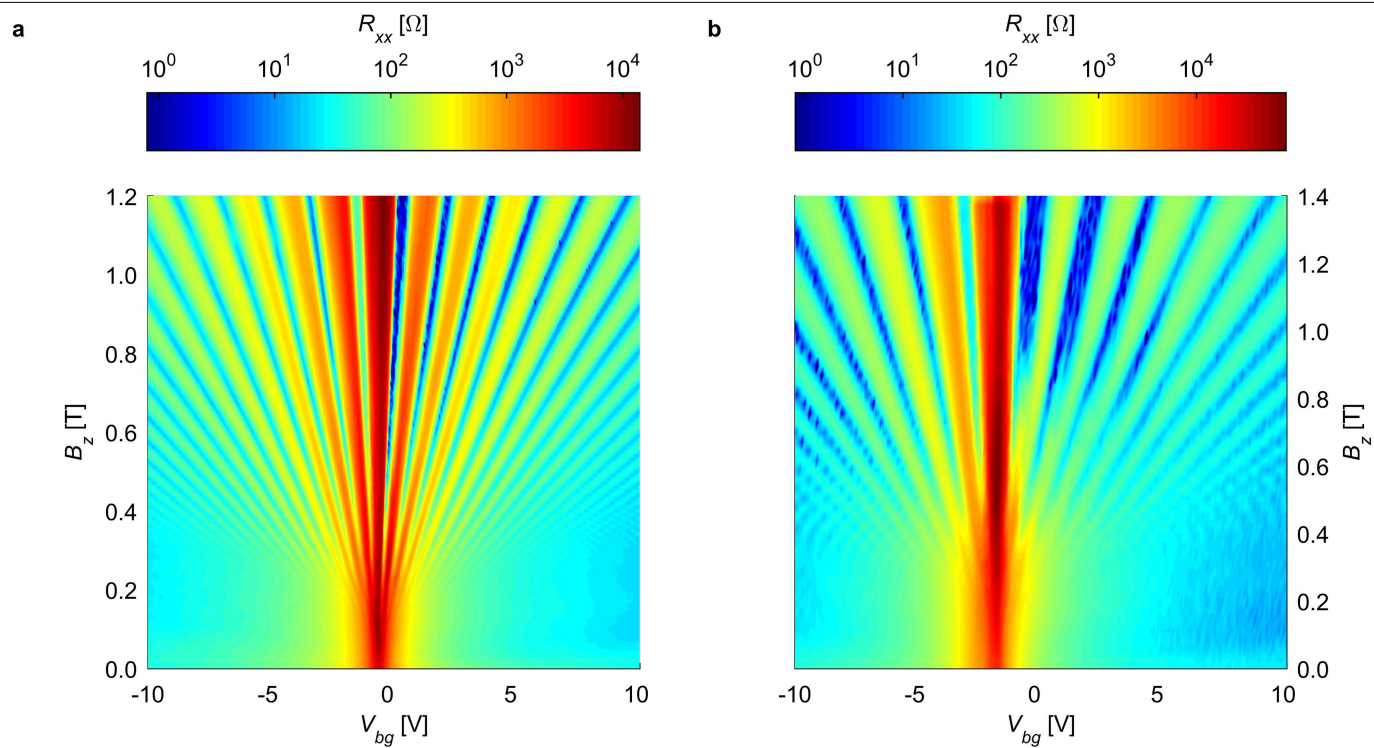
Extended Data Fig. 2 | Optical and scanning electron microscopy images of sample B. **a.** Optical image showing the hBN/graphene/hBN heterostructure (light green), etched trenches exposing the SiO₂/Si substrate (light blue), bottom plunger gates (light brown) and the metal contacts (dark). The dashed rectangle marks the region shown in Fig. 3, Extended Data Figs. 6, 7, 9, Supplementary Fig. 2 and with variable voltage V_{pg} applied to the plunger gates.

The current is applied to the top contact and drained at the bottom contact and the corresponding voltages V_{xx} and V_{2p} are measured in the scanning gate mode. **b.** Scanning electron micrograph of a twin sample of device B showing (from bright to dark) the metal contacts, four plunger gates, hBN/graphene/hBN, and the etched trenches.

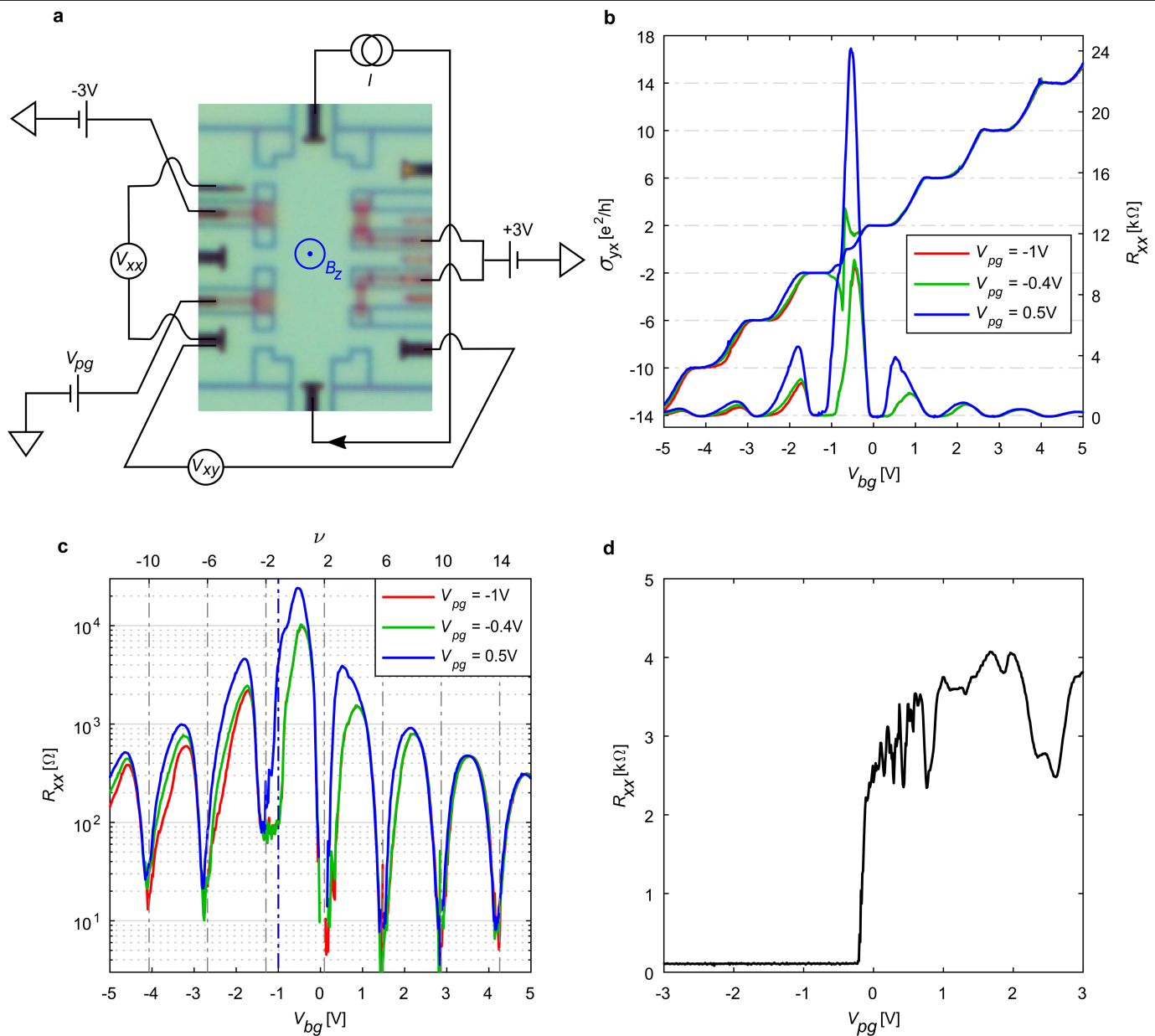


Extended Data Fig. 3 | Optical image of sample C. Shown are the hBN/graphene/hBN heterostructure (light brown), etched trenches exposing the SiO_2/Si substrate (light cyan) and the metal contacts (yellow). The bottom plunger gates are difficult to distinguish in the optical image and are artificially

highlighted in a dark orange colour. The dashed rectangle marks the region shown in Supplementary Fig. 2. The current is applied to the top contact and drained at the bottom, and the corresponding voltages V_{xx} and V_{2p} are measured in the scanning gate mode.



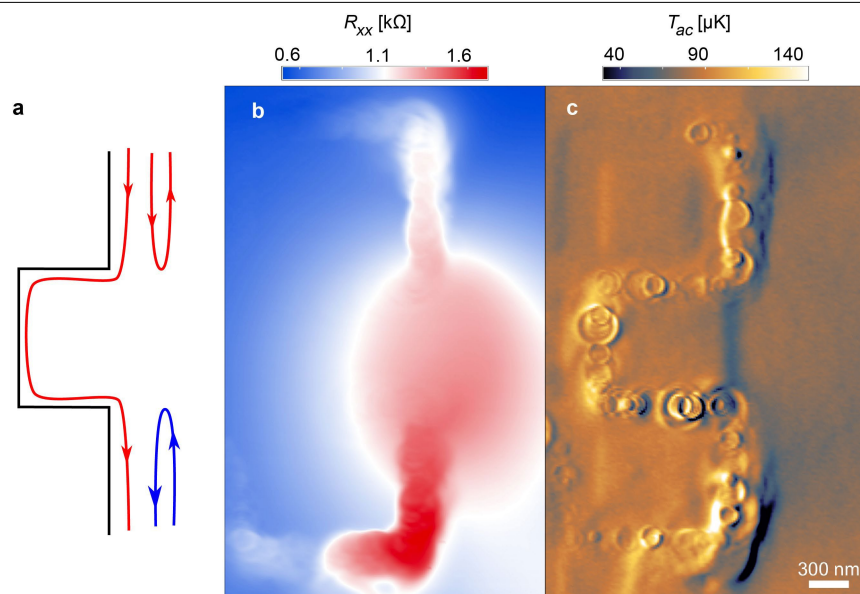
Extended Data Fig. 4 | Transport measurements of samples B and C. a, b, Colour rendering of R_{xx} of samples B (a) and C (b) as a function of the back-gate voltage V_{bg} and the applied perpendicular magnetic field B_z .



Extended Data Fig. 5 | Effect of the plunger gate on transport

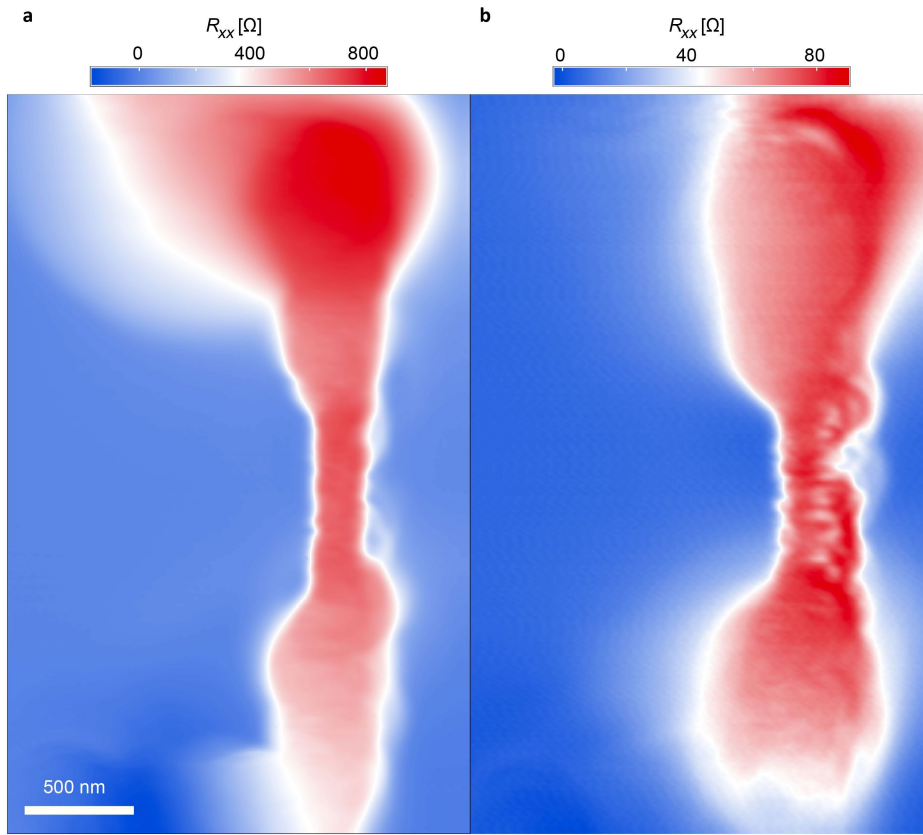
characteristics. a, An optical image of device B with the measurement circuit shown. **b**, Four-probe measurements of Hall conductance $\sigma_{yx} = R_{yx}/(R_{xx}^2 + R_{yx}^2)$ and R_{xx} against back-gate voltage V_{bg} for different V_{pg} values at $B_z = 0.9$ T and $I_{ac} = 50$ nA at 93.72 Hz. A voltage of 3 V was applied to the two plunger gates on the right edge and -3 V was applied to the top plunger gate on the left edge. In this

configuration the nontopological quantum Hall channels on the right edge of the sample are cut off, giving rise to enhanced R_{xx} response on the left edge upon varying V_{pg} . **c**, R_{xx} plotted against V_{bg} from **b** plotted on a logarithmic scale. **d**, Values from the four-probe measurement of R_{xx} plotted against V_{pg} at $V_{bg} = -1$ V (dashed line in c).



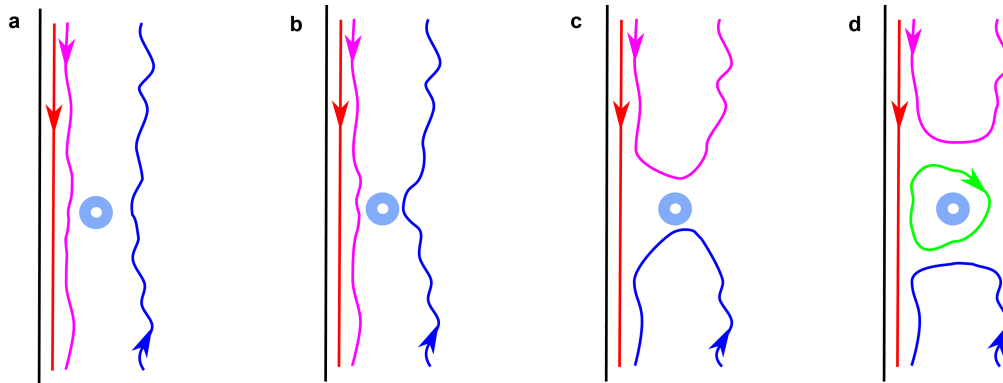
Extended Data Fig. 6 | Visualization of bulk current flow upon cutting off nontopological channels. **a**, Schematic trajectories of quantum Hall edge channels with the nontopological pair of channels cut off by the hole-depleting plunger gate. **b**, Scanning gate $R_{xx}(\mathbf{r})$ image in sample B at $V_{bg} = -1.2$ V, $V_{tg} = 3$ V, $V_{pg} = -0.1$ V and $I_{dc} = 1.75$ μ A, revealing current flowing through the bulk the in

the cut-off region (diffuse red blob). **c**, $T_{ac}(\mathbf{r})$ acquired simultaneously with $R_{xx}(\mathbf{r})$ showing Q rings along the graphene boundaries due to nonlocal dissipation. The images were acquired in the dashed red area in Extended Data Fig. 2.



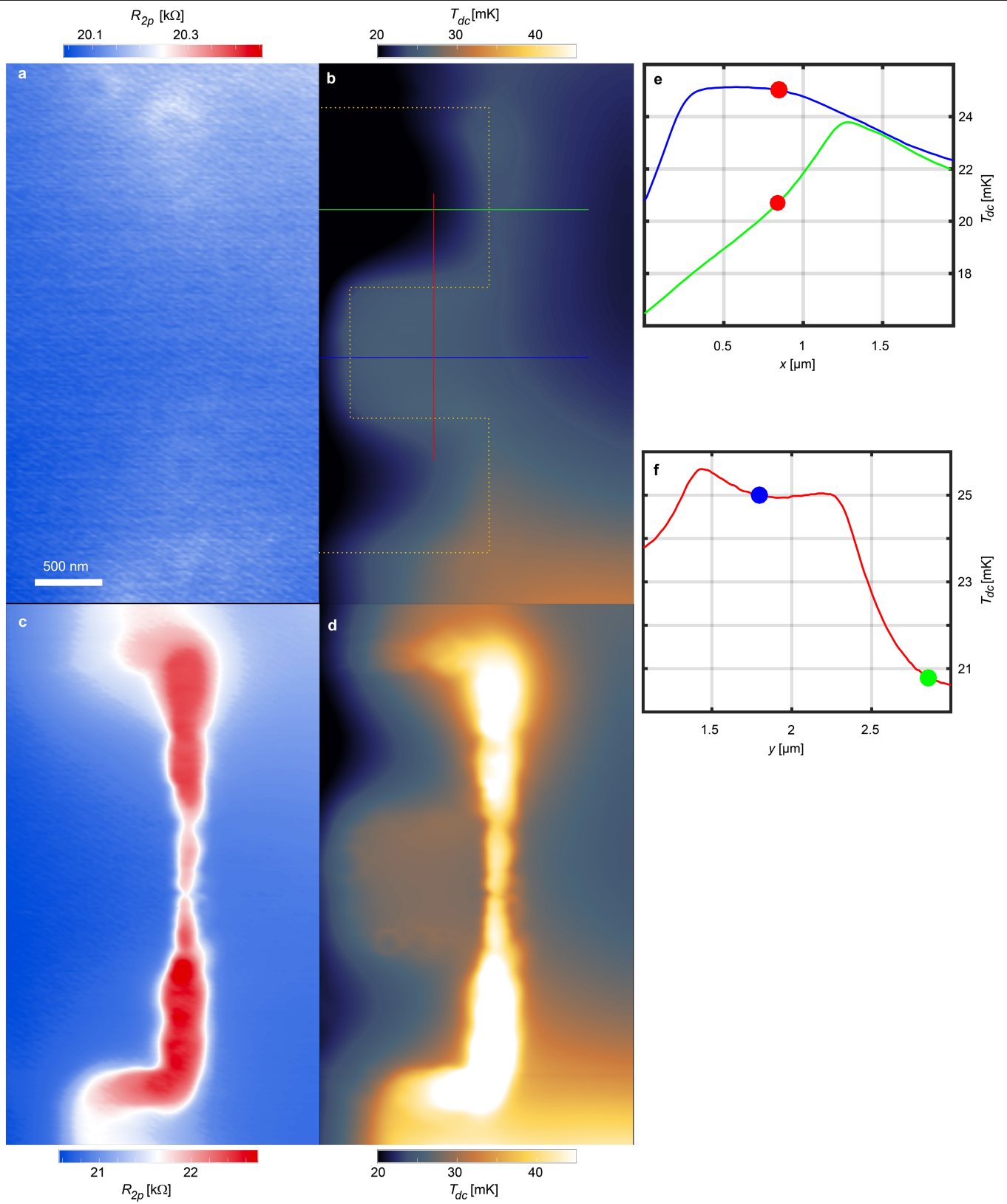
Extended Data Fig. 7 | Demonstration of hole edge accumulation in n-doped bulk. a, Scanning gate $R_{xx}(\mathbf{r})$ imaging of sample B (in the red dashed area in Extended Data Fig. 2) in the vicinity of n-doped $\nu=2$ plateau ($V_{bg}=0.12$ V, $\nu=2.07$, $V_{pg}=-2$ V, see Extended Data Fig. 5c for transport) using a positive V_{tg} of 6 V.

The depletion of the hole accumulated edges by the positive V_{tg} cases increase in $R_{xx}(\mathbf{r})$ similar to the case of the p-doped bulk. **b,** Same as **a** in the vicinity of the $\nu=6$ plateau ($V_{bg}=1.475$ V, $\nu=5.98$). In both images, $I_{dc}=1.75$ μ A.



Extended Data Fig. 8 | Effect of the tip potential V_{tg} on the quantum Hall channels with p-doped edge accumulation. a–d, Schematic trajectories of the edge channels upon increasing V_{tg} . **a**, Non-perturbing tip at flat band conditions $V_{\text{tg}} = V_{\text{tg}}^{\text{FB}} \approx 0\text{V}$. **b**, Application of a weakly perturbing V_{tg} slightly

reduces the edge hole accumulation and shifts the nontopological quantum Hall channels closer to each other. **c**, A stronger depleting V_{tg} cuts off the nontopological pair of channels. **d**, A higher V_{tg} forms an n-doped region under the tip.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Demonstration of elastic tunnelling by comparing perturbing and non-perturbing tip potential in sample B. **a**, Two probe $R_{2p}(\mathbf{r})$ in the case of non-perturbing $V_{tg} = 0.05V \approx V_{tg}^{FB}$ showing essentially constant $R_{2p}(\mathbf{r})$. **b**, The corresponding $T_{dc}(\mathbf{r})$ shows the current-induced temperature variation in the sample unperturbed by the tip at $V_{bg} = -1.1V$ ($v = -1.44$), $V_{pg} = -2V$ and $I_{dc} = 1.75 \mu A$. The increased temperature at the bottom-right corner is caused by heat diffusion from the hot spot at the nearby current contact. **c**, $R_{2p}(\mathbf{r})$ for $V_{tg} = 3V$ revealing the location of $\dot{W}(\mathbf{r})$ processes by perturbing the local work by $\delta\dot{W}(\mathbf{r})$ through enhanced backscattering. **d**, The corresponding

$T_{dc}(\mathbf{r})$ showing the temperature map mimicking the $R_{2p}(\mathbf{r})$ signal caused by the enhanced nonlocal heat release \dot{Q} due to tip-induced $\delta\dot{W}(\mathbf{r})$. **e**, Horizontal line cuts of $T_{dc}(\mathbf{r})$ along the green and blue lines in **b**. The green data show a peak at the graphene boundary (dashed yellow line in **b**) followed by a slowly decaying tail into the bulk, whereas the blue data display no peak at the inner edge of the plunger gates, showing that the $\dot{W}(\mathbf{r})$ process there is elastic. **f**, Vertical line cut through the protrusion region showing peaks at the graphene boundaries with overlapping tails in the middle. The coloured dots are the intersection points of the lines.

Highly efficient and stable InP/ZnSe/ZnS quantum dot light-emitting diodes

<https://doi.org/10.1038/s41586-019-1771-5>

Received: 11 May 2019

Accepted: 8 October 2019

Published online: 27 November 2019

Yu-Ho Won¹, Oul Cho¹, Taehyung Kim¹, Dae-Young Chung¹, Taehee Kim², Heejae Chung¹, Hyosook Jang¹, Junho Lee¹, Dongho Kim² & Eunjoo Jang^{1*}

Quantum dot (QD) light-emitting diodes (LEDs) are ideal for large-panel displays because of their excellent efficiency, colour purity, reliability and cost-effective fabrication^{1–4}. Intensive efforts have produced red-, green- and blue-emitting QD-LEDs with efficiencies of 20.5 per cent⁴, 21.0 per cent⁵ and 19.8 per cent⁶, respectively, but it is still desirable to improve the operating stability of the devices and to replace their toxic cadmium composition with a more environmentally benign alternative. The performance of indium phosphide (InP)-based materials and devices has remained far behind those of their Cd-containing counterparts. Here we present a synthetic method of preparing a uniform InP core and a highly symmetrical core/shell QD with a quantum yield of approximately 100 per cent. In particular, we add hydrofluoric acid to etch out the oxidative InP core surface during the growth of the initial ZnSe shell and then we enable high-temperature ZnSe growth at 340 degrees Celsius. The engineered shell thickness suppresses energy transfer and Auger recombination in order to maintain high luminescence efficiency, and the initial surface ligand is replaced with a shorter one for better charge injection. The optimized InP/ZnSe/ZnS QD-LEDs showed a theoretical maximum external quantum efficiency of 21.4 per cent, a maximum brightness of 100,000 candelas per square metre and an extremely long lifetime of a million hours at 100 candelas per square metre, representing a performance comparable to that of state-of-the-art Cd-containing QD-LEDs. These as-prepared InP-based QD-LEDs could soon be usable in commercial displays.

Intensive efforts to develop QD-LEDs as next-generation displays^{1–6} have shown that their external quantum efficiency (EQE) can be improved up to the theoretical maximum (20.5%)⁴ by optimizing the gradient core/shell structures of the QDs^{7–9} and by adopting an inorganic electron transport layer² with electron–hole blocking layers⁴. However, most previous work has focused on CdSe-based QDs, which present severe toxicity and environmental issues. Few studies have investigated the more environmentally benign InP-based QD-LEDs, because of the difficulty in synthesizing high-quality materials (Extended Data Table 1). Recently, InP/ZnSe/ZnS QDs prepared through precursor purification have been shown to have a high quantum yield of 93% but the corresponding QD-LED showed an EQE of 12.2% without stability data¹⁰. The poor performance of the InP-based QD-LEDs was attributed to defects in the deep in-gap states of InP QDs^{11,12} and oxidative defects¹³. Here we describe a way to prepare InP/ZnSe/ZnS QDs with excellent characteristics: a perfect quantum yield and the narrowest reported (to our knowledge) full-width at half-maximum (FWHM; 35 nm at 630 nm) with a highly spherical shape. The uniformity of the InP core is greatly improved by adding two consecutive steps: nucleation and controlled growth through continuous precursor injection. For shell passivation, the primary ZnSe interlayer on the core is formed simultaneously with in situ etching of the oxide surface of the InP using hydrofluoric acid

(HF). In addition, the thickness of the ZnSe interlayer is increased up to 3.6 nm to suppress non-radiative Auger recombination and Förster resonance energy transfer (FRET), which eventually influences the EQE and lifetime of the QD-LEDs. Furthermore, the surface ligands of the QDs are tailored for better charge injection in the QD-LEDs. Our QD-LED achieved the theoretical maximum EQE of 21.4% and a high brightness of 100,000 cd m^{–2}. It also has an outstandingly long operating half-life of 1,000,000 h at 100 cd m^{–2}.

Oxidative defects are suspected to be the major cause of the poor optical properties of InP; hence, many researchers have attempted to reduce oxygen sources^{12,14}. Although oxygen-free precursors have been used, the spectra obtained were broad (around 46–63 nm), and the quantum yields were low (around 20–60%), owing to oxygen contamination arising from impurities generated during the reaction¹². We therefore thoroughly purified the precursors and solvent under a vacuum of 1 mTorr at 120 °C before the reaction and N₂ flow during the synthesis. To form uniform InP cores, two consecutive steps were applied: first, initial nucleation and growth and then further growth using additional In/P precursors. Figure 1a illustrates the synthesis of InP/ZnSe/ZnS QDs with different shell structures. The first absorption maximum (*A*₁) after the injection of the P precursor appeared at 500 nm, indicating a diameter of 2.7 nm (core 1). The bandgap was further tuned

¹Samsung Advanced Institute of Technology, Samsung Electronics, Suwon, South Korea. ²Department of Chemistry, Yonsei University, Seoul, South Korea. *e-mail: ejjang12@samsung.com

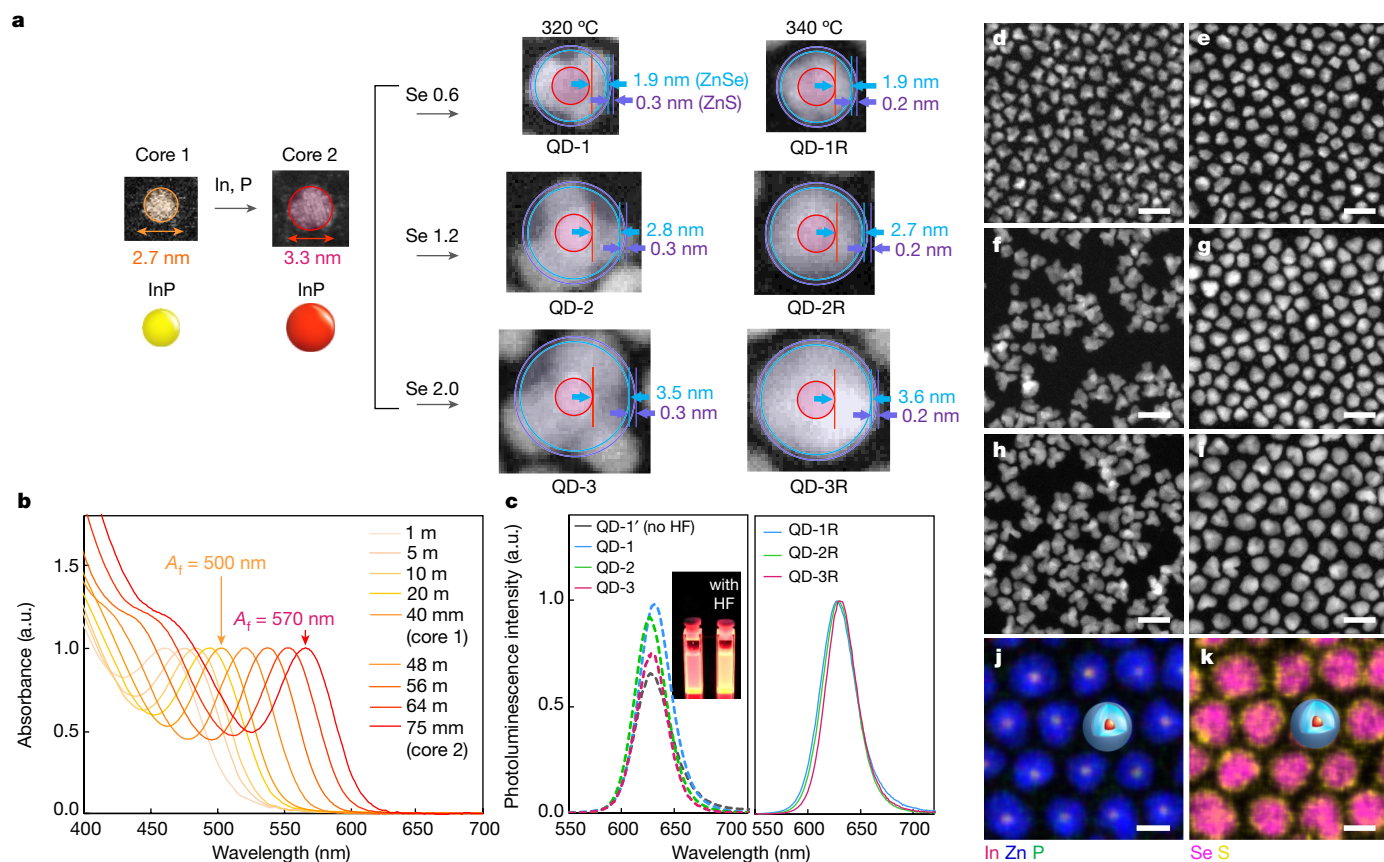


Fig. 1 | InP/ZnSe/ZnS QDs with different morphology and shell thickness.

a, Preparation of InP cores and InP/ZnSe/ZnS QDs with different morphology and shell thickness. The amounts of Se precursor for QD-1, QD-2 and QD-3 were 0.6 mmol, 1.2 mmol and 2.0 mmol, respectively, per 10 ml of solvent. The estimated size, based on inductively coupled plasma-atomic emission spectroscopy (ICP-AES) data, was projected on the STEM image of each QD.

b, Ultraviolet–visible absorption spectra of the aliquots, taken during the InP

core synthesis. a.u., arbitrary units. **c**, Photoluminescence spectra of QD-1' (prepared without HF addition), QD-1, QD-2, QD-3, QD-1R, QD-2R and QD-3R. Inset, photograph of QD-1' (no HF) and QD-3 taken under 365 nm illumination.

d–i, STEM images of QD-1, QD-1R, QD-2, QD-2R, QD-3 and QD-3R (scale bar, 20 nm). **j, k**, Electron diffraction spectroscopy mapping of In, Zn, P, Se and S for QD-3R (scale bar, 10 nm).

on adding In/P precursors: a larger core (core 2) with an A_f at 570 nm, indicating a diameter of 3.3 nm, was produced (Fig. 1b). Moreover, the sharpness of the absorption peak of the core had a linear correlation with the FWHM in the photoluminescence of the core/shell (Extended Data Fig. 1a, b). The X-ray diffraction pattern of core 2 was consistent with the structure of zinc-blende, and its size and shape distributions were confirmed as homogeneous by scanning transmission electron microscopy (STEM) (Extended Data Fig. 1d, e). As both In and P are oxophilic, defects such as InPO_x or In_2O_3 can easily be generated during synthesis^{11,13}. HF treatment has proved to be effective for removing oxides¹⁵; however, shell coating after the treatment resulted only in poor passivation¹⁶. Here, we added the HF solution at an early stage of shell growth, and the ZnSe shell was grown simultaneously by injecting a Se precursor with HF to prevent re-oxidation. The ZnS shell was then coated on the InP/ZnSe QDs by adding Zn and S precursors sequentially. The growth of the ZnSe (1.9 nm) and ZnS shell (0.3 nm) produced red-emitting InP/ZnSe/ZnS QDs (QD-1) with a quantum yield of 98%—a great improvement over that of QDs prepared without HF (80%) (Fig. 1c, Extended Data Table 2). X-ray photoelectron spectroscopy confirmed that the InPO_x peak at 132 eV in the P_{2p} spectrum disappeared on adding HF to the InP core (Extended Data Fig. 1c). However, lattice mismatch between InP and ZnSe induced a pod-shaped evolution during shell growth, implying non-uniform passivation. As we increased the thickness of ZnSe (QD-2 and QD-3), the quantum yields decreased (to 92% and 75%) and the shapes obtained became more irregular (Fig. 1f, h). To prepare more spherical InP/ZnSe/ZnS with uniform shells, we raised

the reaction temperature to 340 °C to facilitate a kinetically controlled reaction on a random facet (QD-1R, QD-2R and QD-3R)¹⁷. On assigning the degree of shell evolution, assuming a spherical shape based on the elemental compositions obtained by inductively coupled plasma spectroscopy (Fig. 1a), we noted a discrepancy with the sizes obtained by transmission electron microscopy (TEM), due to the shape difference. However, the size of the spherical QDs was in accordance with the TEM result. The photoluminescence peak and FWHM of QD-3R were 630 nm and 35 nm, respectively, and the quantum yield was enhanced up to 100%. The quantum yields of QD-1R and QD-2R also increased to 100% (Fig. 1c). X-ray diffraction patterns of the QDs agreed well with those of the ZnSe bulk reference (zinc-blende), and the HAADF-STEM image of QD-3R projected along the $[111]$ axis showed a zinc-blende crystal structure (Extended Data Fig. 2). Energy-dispersive X-ray spectroscopy images of QD-3R clearly revealed a InP/ZnSe/ZnS core/multi-shell structure (Fig. 1j, k).

We applied differently structured QDs in LED devices comprising ITO/PEDOT:PSS (35 nm)/TFB (25 nm)/QD (20 nm)/ZnMgO (40 nm)/Al (100 nm)⁹ (see Methods for the fabrication and characteristics of the QD-LEDs; Fig. 2a). The turn-on voltages of all QD-LEDs were about 1.8–2.0 V, close to the optical bandgap of InP QDs (1.97 eV). Their current and voltage profiles were almost identical, demonstrating similar charge transport behaviour (Fig. 2b). This resemblance of the current–voltage character of the QD-LEDs despite the different shell thicknesses was also confirmed by the ionization potential of the QDs and the current–voltage profiles of the single-carrier devices (Extended

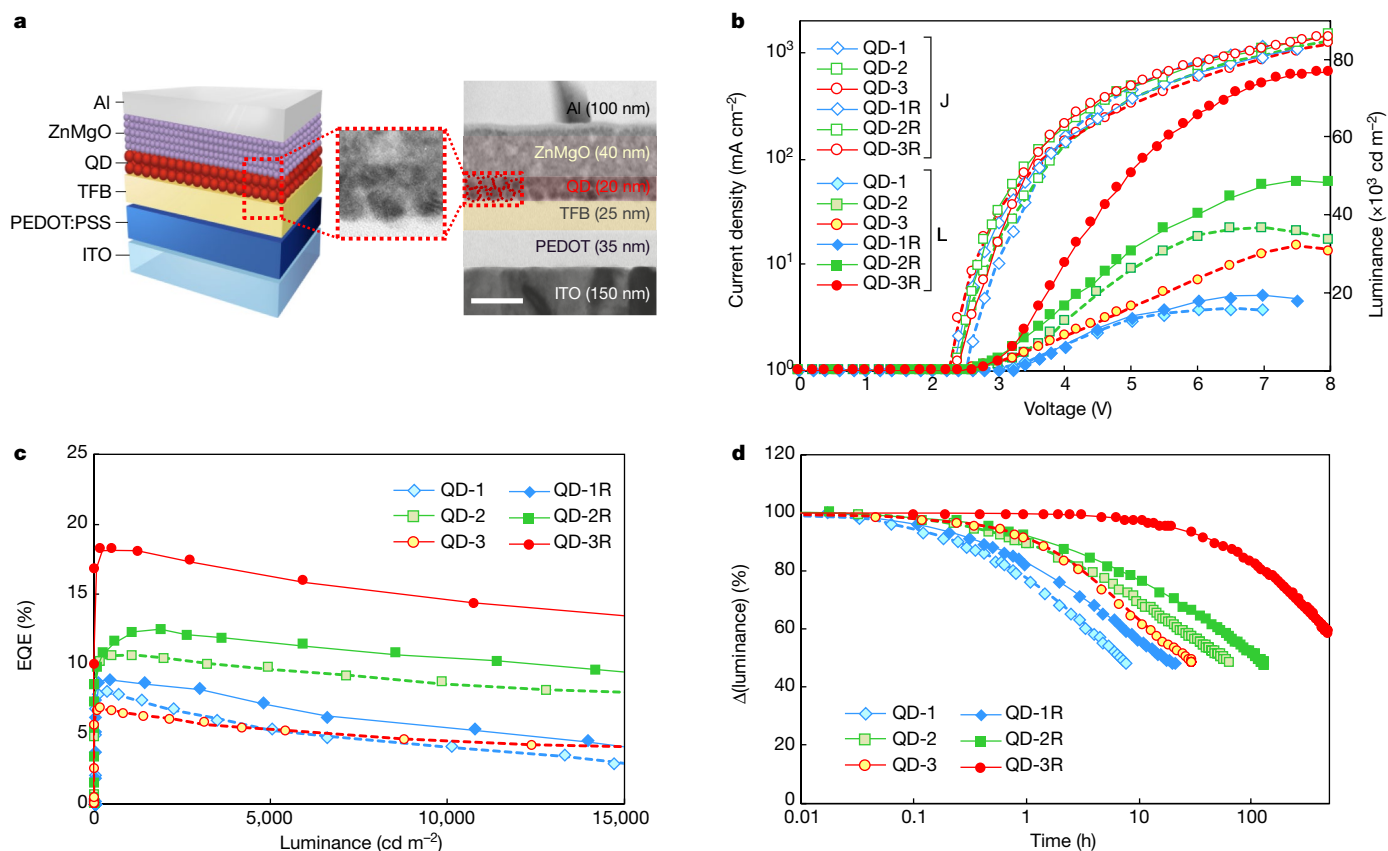


Fig. 2 | Performances of InP-based QD-LEDs. **a**, Illustration and cross-sectional TEM image of the QD-LED device structure (scale bar, 50 nm). **b**, Current density (left axis) and luminance (right axis) versus voltage profiles. **c**, EQE–

luminance profiles. **d**, Lifetime measurements (at the initial luminance of 4,500 cd m⁻²) of the QD-LEDs with QD-1, QD-1R, QD-2, QD-2R, QD-3 and QD-3R.

Data Fig. 3a–c). The similarity of the valence band maxima for all QDs (−5.6 eV) irrespective of the ZnSe thickness could be due to their having the same InP core. Furthermore, electron–hole transports were similar for all single-carrier devices. We postulate that the total current density is more dependent on the electron transport, which can be affected by the shell thickness and the number of the QD layers. However, the EQEs and luminance of the LEDs were greatly enhanced on increasing the ZnSe interlayer thickness. The more spherical QDs had slightly improved device performances owing to their increased photoluminescence quantum yields. QD-LEDs with QD-3R showed an EQE of 18.0% and a maximum luminance (L_{\max}) of 70,718 cd m⁻², which exceeded those of other QD-LEDs (Fig. 2c). The QD-LED with QD-3 showed very poor EQE and luminance owing to its low photoluminescence quantum yield (75%). The lifetimes of the devices were strongly affected by the ZnSe shell thickness. The time taken for a 25% drop in the brightness (T_{75}) at the initial luminance of 4,500 cd m⁻² for the LEDs was affected most by the shell thickness (Fig. 2d; the T_{75} values for QD-1, QD-1R, QD-2, QD-2R, QD-3 and QD-3R were 1.3 h, 2.3 h, 6.2 h, 12.6 h, 4.7 h and 200 h, respectively). The QD-LED with QD-3R showed a superior lifetime, and its T_{75} at 1,000 cd m⁻² is predicted to be 3,000 h, using the acceleration factor of 1.8 (given that we measured the real lifetime at 4,500 cd m⁻²).

To understand why device performance improved with increasing shell thickness, we performed spectroscopic analysis. For Cd-based QDs, a thick shell¹⁸ or an alloy shell¹⁹ was beneficial in suppressing Auger recombination, which causes QD charging and degradation of device efficiency. A thick shell can increase the inter-particle distance and reduce energy transfer among neighbouring QDs²⁰. The static and time-resolved photoluminescence spectra reconstructed from transient

decay for the 0–20 ns regime are shown in Fig. 3a and Extended Data Fig. 4a–c. The photoluminescence of the QD film red-shifted compared to that of the solution, and the degree of shift decreased as the ZnSe shell thickness increased regardless of the morphology, indicating that a thick shell can suppress FRET effectively. Additionally, the photoluminescence intensity of the QD film declined less as the ZnSe shell thickness increased. The relative photoluminescence intensities at 20 ns with respect to 0 ns were 35%, 40% and 56% for QD-1, QD-2 and QD-3R, respectively; the corresponding values for the solutions were 56%, 57% and 63%. This indicates that FRET influences the photoluminescence efficiency in a close-packed QD film. The photoluminescence quantum yields for the QD-1, QD-2 and QD-3R films obtained from the integrating sphere were 69%, 84% and 93%, respectively, indicating better preservation of quantum yield with increasing shell thickness. This agrees with previously reported results for Cd-based QDs¹⁸; nevertheless, the effective shell thickness could differ depending on the band alignments of the core and shell materials. We calculated the spatial distribution of the electron–hole wave functions for the QDs using effective mass approximation (Extended Data Fig. 4j–l), demonstrating that the hole was confined around the InP core (2 nm) and that the electron was delocalized to 4 nm from the QD radius. From effective mass approximation calculations and structural information, we calculated the FRET efficiency (E_f) for the QDs with different shell thicknesses according to the previous literature²¹ except that we considered the dipole interactions in the effective core volume (Extended Data Table 3). E_f decreased as the shell thickness increased (46%, 15% and 6% for QD-1, QD-2 and QD-3R, respectively), implying suppressed photoluminescence quenching for the thick-shelled QDs in the film. We

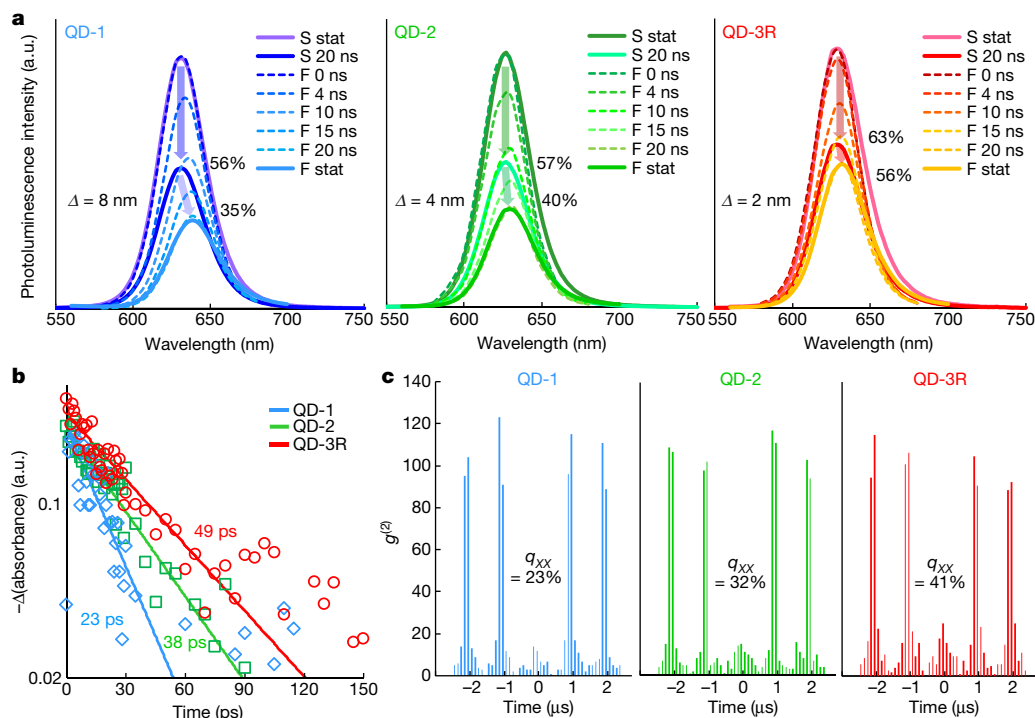


Fig. 3 | Optical characteristics of QD-1, QD-2 and QD-3R. a, Static photoluminescence spectra of the QD solution (S) and the QD film (F) ('stat', solid lines) together with dynamic photoluminescence spectra (dashed lines) of the QD film collected at each decay time (0–20 ns). **b**, Traces of Auger decay

extracted from pump-fluence-dependent transient absorption dynamics. The pumping intensity used in the experiments was 0.4 mW for QD-1 and QD-2 and 0.32 mW for QD-3R. **c**, $g^{(2)}$ statistics and extracted biexciton quantum yields (q_{xx}) for QD-1, QD-2 and QD-3R single dots.

also evaluated E_f from the time-resolved photoluminescence measurements of QDs in solution and film (51%, 28% and 10% for QD-1, QD-2 and QD-3R, respectively in Extended Data Fig. 4d–i), which agrees with the theoretical calculations. Moreover, the unbalanced carrier injection

in the LED devices could induce the generation of multiple excitons in the QDs, causing Auger recombination. To analyse the Auger effect, transient absorptions were recorded from a single-exciton regime to a multi-exciton one by controlling the excitation intensity (Extended

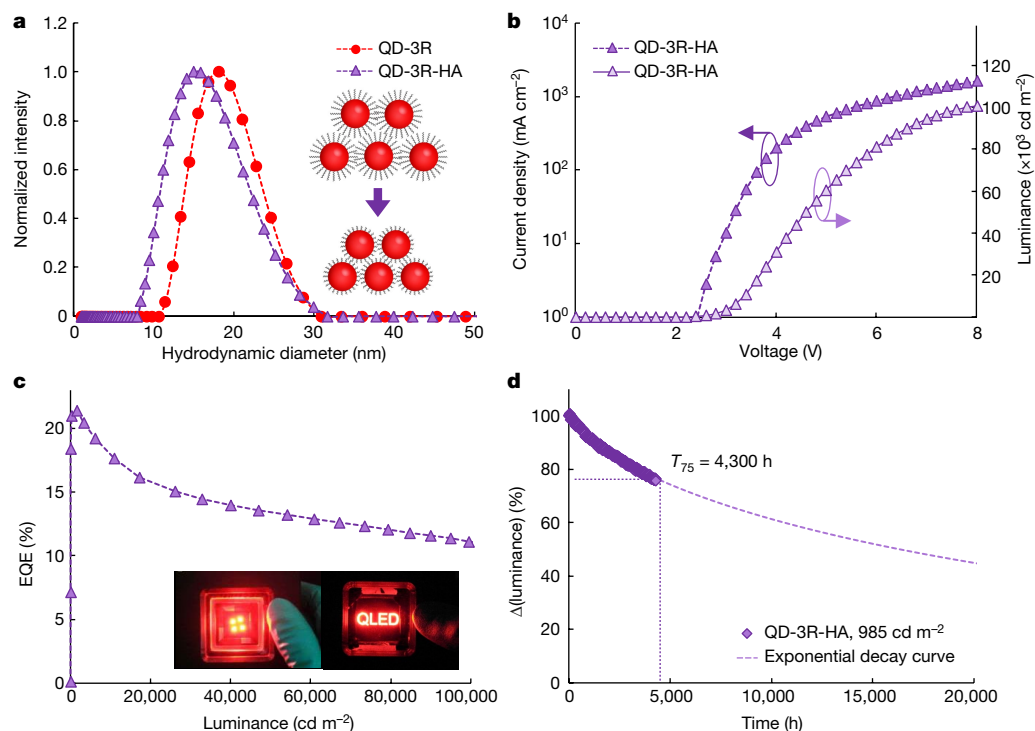


Fig. 4 | QD-LED with ligand-exchanged QD. a, Dynamic laser scattering spectra of QD-3R and QD-3R-HA. **b**, Current density (left axis) and luminance (right axis) versus voltage profiles. **c**, EQE–luminance profile. Inset,

photographs of four-pixel QD-LED and text-patterned QD-LED. **d**, Lifetime measurement (at the initial luminance of 985 cd m^{-2}) of the QD-LED with QD-3R-HA fitted with exponential decay curve, $y = 100 \exp(-0.00079x^{0.7})$.

Data Fig. 4m–o). Under intense excitation, where the average number of excitons per QD is larger than one, the bleaching dynamics of the first excitonic transition showed characteristic fast decay due to non-radiative Auger recombination²². Figure 3b shows that the extracted Auger lifetime increases as the ZnSe shell thickness increases (23 ps, 38 ps and 49 ps for QD-1, QD-2 and QD-3R, respectively), indicating more effective Auger suppression in the thicker ZnSe shell. These results can be explained by a recent theoretical study²³ according to which the Auger rate was mostly dominated by the positive trion channel suppressed by thick-shell passivation wherein the hole was more confined than the electron. The Auger suppression should accompany an improvement in the biexciton emission efficiency, which can be evaluated from the second-order intensity correlation functions²⁴ of a single dot. Figure 3c shows that the relative biexciton quantum yield indeed increases with increasing ZnSe thickness (23%, 32% and 41% for QD-1, QD-2 and QD-3R, respectively).

Charge transport in the QD-LEDs is also influenced by surface ligands. Our QDs had an oleic acid (OA) ligand, which acted as an effective barrier for electron or hole carriers, and hexanoic acid (HA) to reduce the alkyl length. After ligand exchange, 63% of the OA was replaced with HA (QD-3R-HA), as confirmed by pyrolysis gas chromatography–mass spectrometry; further, the solid residue increased from 87 to 92 weight per cent, maintaining the total amount of ligands, according to thermogravimetric analysis (Extended Data Fig. 5). The hydrodynamic size, measured by dynamic light scattering, decreased from 18.5 nm to 16.1 nm (Fig. 4a). Thick-shelled QDs with short ligands (QD-3R-HA) were then applied to a LED, which exhibited an EQE of 21.4% and a L_{max} of 100,000 cd m^{-2} (Fig. 4b–d). This EQE was comparable to the theoretical maximum and was the highest reported for QD-LEDs thus far. Moreover, we fabricated 47 devices to confirm the reproducibility; they showed EQE values of $19.8\% \pm 1.1\%$ and turn-on voltages of approximately 1.8 V (Extended Data Fig. 6). More importantly, the QD-LED with QD-3R-HA exhibited an excellent lifetime ($T_{95} = 615$ h at $1,000 \text{ cd m}^{-2}$). The measured T_{75} lifetime at $1,000 \text{ cd m}^{-2}$ was 4,300 h (Fig. 4d), which was longer than that for QD-3R (3,000 h); moreover, T_{50} was extrapolated as 16,348 h using an experimental fitting curve, $y = 100\exp(-0.00079x^{0.7})$. The predicted T_{50} at 100 cd m^{-2} was 1,000,000 h, based on the T_{50} at $1,000 \text{ cd m}^{-2}$ with an acceleration factor of 1.8 (Extended Data Fig. 7). To our knowledge, this is the best lifetime reported for InP-based QD-LEDs and is comparable to the best performance noted for Cd-based QD-LEDs⁹. To investigate the effect of the ligand on the charge transport of the QDs, single-carrier devices were also fabricated (Extended Data Fig. 6c). The hole current increased up to fourfold for the short ligand, causing exciton recombination and reducing Auger recombination, thereby suppressing the voltage increase at the interface between the hole transport layer (TFB in this case) and the QD during operation (Extended Data Fig. 6d).

In summary, we present excellent InP/ZnSe/ZnS QDs showing a perfect quantum yield (100%), narrow spectrum (FWHM 35 nm) and highly symmetric shape. Moreover, the devices using them exhibit excellent values for the EQE (21.4%), brightness (100,000 cd m^{-2}), and outstanding lifetime (1,000,000 h at 100 cd m^{-2}). These properties are highly superior to those of previously reported Cd-free QD-LEDs and comparable to those of state-of-the-art Cd-based QD-LEDs. Hence, the results will aid in fabricating Cd-free QD-LEDs for next-generation displays.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1771-5>.

- Coe, S., Woo, W.-K., Bawendi, M. & Bulovic, V. Electroluminescence from single monolayers of nanocrystals in molecular organic devices. *Nature* **420**, 800–803 (2002).
- Qian, L., Zheng, Y., Xue, J. & Holloway, P. Stable and efficient quantum-dot light-emitting diodes based on solution-processed multilayer structures. *Nat. Photon.* **5**, 543–548 (2011).
- Mashford, B. S. et al. High-efficiency quantum-dot light-emitting devices with enhanced charge injection. *Nat. Photon.* **7**, 407–412 (2013).
- Dai, X. et al. Solution-processed, high-performance light-emitting diodes based on quantum dots. *Nature* **515**, 96–99 (2014).
- Manders, J. R. et al. High efficiency and ultra-wide color gamut quantum dot LEDs for next generation displays. *J. Soc. Inf. Disp.* **23**, 523–528 (2015).
- Wang, L. et al. Blue quantum dot light-emitting diodes with high electroluminescent efficiency. *ACS Appl. Mater. Interf.* **9**, 38755–38760 (2017).
- Bae, W. K. et al. Controlling the influence of Auger recombination on the performance of quantum-dot light-emitting diodes. *Nat. Commun.* **4**, 2661 (2013).
- Yang, Y. et al. High-efficiency light-emitting devices based on quantum dots with tailored nanostructures. *Nat. Photon.* **9**, 259–266 (2015).
- Cao, W. et al. Highly stable QLEDs with improved hole injection via quantum dot structure tailoring. *Nat. Commun.* **9**, 2608 (2018).
- Li, Y. et al. Stoichiometry-controlled InP-based quantum dots: synthesis, photoluminescence, and electroluminescence. *J. Am. Chem. Soc.* **141**, 6448–6452 (2019).
- Reiss, P., Carriere, M., Lincheneau, C., Vaure, L. & Tamang, S. Synthesis of semiconductor nanocrystals, focusing on nontoxic and earth-abundant materials. *Chem. Rev.* **116**, 10731–10819 (2016).
- Tessier, M. D., Dupont, D., Nolf, K. D., Roo, J. D. & Hens, Z. Economic and size-tunable synthesis of InP/ZnS (E=Se, S) colloidal quantum dots. *Chem. Mater.* **27**, 4893–4898 (2015).
- Stein, J. L. et al. Probing surface defects of InP quantum dots using phosphorus K α and K β X-ray emission spectroscopy. *Chem. Mater.* **30**, 6377–6388 (2018).
- Tessier, M. D. et al. Interfacial oxidation and photoluminescence of InP-based core/shell quantum dots. *Chem. Mater.* **30**, 6877–6883 (2018).
- Mićić, O. I., Jones, K. M., Cahill, A. & Nozik, A. J. Optical, electronic, and structural properties of uncoupled and close-packed arrays of InP quantum dots. *J. Phys. Chem. B* **102**, 9791–9796 (1998).
- Mićić, O. I., Smith, B. B. & Nozik, A. J. Core-shell quantum dots of lattice-matched ZnCdSe₂ shells on InP cores: experiment and theory. *J. Phys. Chem. B* **104**, 12149–12156 (2000).
- Liu, L. et al. Shape control of CdSe nanocrystals with zinc blende structure. *J. Am. Chem. Soc.* **131**, 16423–16429 (2009).
- Lim, J. et al. Influence of shell thickness on the performance of light-emitting devices based on CdSe/Zn_{1-x}Cd_xS core/shell heterostructured quantum dots. *Adv. Mater.* **26**, 8034–8040 (2014).
- Bae, W. K. et al. Controlled alloying of the core-shell interface in CdSe/CdS quantum dots for suppression of Auger recombination. *ACS Nano* **7**, 3411–3419 (2013).
- Lakovic, J. R. *Principles of Fluorescence Spectroscopy* (Springer, 2006).
- Kagan, C. R., Murray, C. B. & Bawendi, M. G. Long-range resonance transfer of electronic excitations in close-packed CdSe quantum-dot solids. *Phys. Rev. B* **54**, 8633 (1996).
- Klimov, V. I., Mikhailovsky, A. A., McBranch, D. W., Leatherdale, C. A. & Bawendi, M. G. Quantization of multiparticle Auger rates in semiconductor quantum dots. *Science* **287**, 1011 (2000).
- Vaxenburg, R., Rodina, A., Lifshitz, E. & Efros, A. L. Biexciton Auger recombination in CdSe/CdS core/shell semiconductor nanocrystals. *Nano Lett.* **16**, 2503–2511 (2016).
- Park, Y.-S. et al. Near-unity quantum yields of biexciton emission from CdSe/CdS nanocrystals measured using single-particle spectroscopy. *Phys. Rev. Lett.* **106**, 187401 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Materials

Tris(trimethylsilyl)phosphine (TMS_3P , 95%) and trioctylphosphine (TOP, 97%) were purchased from Strem. Indium acetate ($\text{In}(\text{OAc})_3$, 99.99%), zinc acetate ($\text{Zn}(\text{OAc})_2$, 99.99%), palmitic acid (PA, 99%), oleic acid (OA, 90%), hexanoic acid (HA, 99%), sulfur (S, 99.99%), selenium (Se, 99.99%), hydrofluoric acid (HF, 48%), 1-octadecene (ODE, 90%) and trioctylamine (TOA, 98%), toluene (anhydrous, 99.8%), cyclohexane (anhydrous, 99.5%) and 1-butanol (anhydrous, 99.8%) were purchased from Sigma-Aldrich. Acetone (99.7%) and ethanol (HPLC, 99.9%) were purchased from Samchun Chemicals. PEDOT:PSS in aqueous solution was purchased from H. C. Starck. Poly[(9,9-dioctylfluorenyl-2,7-diyl)-co-(4,4'-(N-(4-sec-butylphenyl)diphenylamine))] (TFB, molecular weight 300,000) and 1,4,5,8,9,11-hexa-azatriphenylene hexacarbonitrile (HATCN) were purchased from Sumitomo Chemical Ltd. and LUMTEC, respectively. Aluminium pellets were purchased from iTASCO.

Preparation of precursor

In general, to prepare 0.2 M indium palmitate, $\text{In}(\text{PA})_3$ precursor, 11 mmol of $\text{In}(\text{OAc})_3$ and 33 mmol of PA were dissolved in 55 ml of ODE, and the mixture was evacuated at 120 °C for 1 h and heated to 280 °C for 10 min under N_2 flow (800 $\text{cm}^3 \text{min}^{-1}$). 0.2 M TMS_3P in TOP was prepared with 5.5 mmol of TMS_3P and 27.5 ml of TOP in an N_2 -filled glove box. 0.4 M $\text{Zn}(\text{OA})_2$ precursor was prepared with 48 mmol of $\text{Zn}(\text{OAc})_2$ and 96 mmol of OA dissolved in 120 ml of TOA, and the mixture was evacuated and heated through the same method as above. For 0.4 M Se/TOP and 1 M S/TOP precursors, Se pellet (7.89 g) and S powder (3.20 g) were dissolved in 100 ml and 40 ml of TOP, respectively, and each solution was evacuated at 120 °C for 30 min inside an N_2 -filled glove box.

Synthesis of InP cores

In a 250 ml flask, 4 mmol of $\text{In}(\text{OAc})_3$ and 12 mmol of PA were dissolved in 100 ml of ODE, and the mixture was evacuated (120 °C, 1 h) and heated (280 °C, 10 min) under N_2 flow. Then, 10 ml of 0.2 M TMS_3P /TOP was quickly injected into the $\text{In}(\text{PA})_3$ solution at 280 °C, which decreased the temperature to 260 °C. The reaction medium was maintained at that temperature for 40 min to prepare core 1 ($A_f = 500 \text{ nm}$). Then, 52.5 ml of 0.2 M $\text{In}(\text{PA})_3$ and 26.3 ml of 0.2 M TMS_3P /TOP was added dropwise at the rate of 1.5 and 0.75 ml min^{-1} for 35 min into the reaction mixture to synthesize core 2 ($A_f = 570 \text{ nm}$). The heating mantle was then removed and the reactor was cooled to room temperature. For shell growth, the InP core was precipitated by adding 80 ml of acetone to 16 ml of the crude reaction mixture, followed by centrifugation at 5,800 rpm for 5 min. The InP core, re-dispersed in toluene (12 ml), showed an optical density of 0.26 at 570 nm (the absorption measured for a 10 μl QD solution in 990 μl toluene).

Synthesis of InP/ZnSe/ZnS QDs

InP/ZnSe/ZnS QDs were prepared by mixing 1.6 mmol of $\text{Zn}(\text{OAc})_2$ and 3.2 mmol of OA in 80 ml of TOA using a 250 ml flask. The mixture was evacuated (120 °C, 1 h) and heated (280 °C, 10 min) under N_2 flow, and then cooled to 180 °C. Then, the InP core in toluene (12 ml, 0.25 g) was injected into the solution at 180 °C. Diluted HF in acetone (10 weight per cent, 0.2 ml) was immediately added after core injection. The quick injection of the solution containing toluene and acetone at 180 °C can rapidly increase the pressure in the reaction flask. To synthesize QD-3R, the solution was heated to 340 °C, following which 17.6 mmol of 0.4 M $\text{Zn}(\text{OA})_2$ and 16 mmol of Se/TOP were added into the reaction mixture and maintained for 1 h to grow the ZnSe shell. The ZnS shell was coated by adding 4.8 mmol of $\text{Zn}(\text{OA})_2$ and 6.4 mmol of S/TOP for 20 min, and the reaction mixture was finally cooled to room temperature. QD-3R was precipitated by adding 30 ml of ethanol to 10 ml

of the crude solution and centrifugation at 5,800 rpm for 5 min. The collected QDs were dispersed in 10 ml of octane and used for further LED fabrication. To prepare QD-3, the amount of the reactants and the procedures, except for the reaction temperature (320 °C) and reaction time (additional 30 min). The amount of each reactant was varied to control the shell thickness: for QD-1 and QD-1R, 11.2 mmol of $\text{Zn}(\text{OAc})_2$ and 4.8 mmol of Se/TOP were used for the ZnSe shell growth, and 4.8 mmol of S/TOP for ZnS, whereas for QD-2 and QD-2R, 12.0 mmol of $\text{Zn}(\text{OAc})_2$, 9.6 mmol of Se/TOP were used for ZnSe, and 4.8 mmol of $\text{Zn}(\text{OAc})_2$ and 5.6 mmol of S/TOP were used for ZnS.

Ligand exchange

The ligand exchange of OA with HA was conducted as follows: $\text{Zn}(\text{HA})_2$ was prepared with 24 mmol of $\text{Zn}(\text{OAc})_2$ and 48 mmol of HA in 120 mL of TOA solvent following the same procedure as that used to prepare the $\text{Zn}(\text{OA})_2$ precursor. The prepared $\text{Zn}(\text{HA})_2$ was characterised with ^1H -nuclear magnetic resonance (NMR) spectra using solutions in toluene- d_8 [2.45 (m, 2H), 1.54 (m, 2H), 1.36–1.21 (m, 4H), 0.93 (t, 3H)], and compared with HA [2.03 (m, 2H), 1.45 (m, 2H), 1.14 (m, 2H), 1.08 (m, 2H), 0.80 (t, 3H)] (Extended Data Fig. 5c).

Finally, 24 mmol of $\text{Zn}(\text{HA})_2$ was mixed with 20 ml of the crude reaction mixture of QD-3R, which was heated to 200 °C and maintained for 30 min. QD-3R-HA was obtained using the same separation method as above.

Material characterization

The absorption and photoluminescence spectra of the QD solutions were measured with an ultraviolet–visible spectrometer (Varian Cary 5000) and fluorescence spectrophotometer (Hitachi F7000), respectively. The photoluminescence quantum yield was obtained as the absolute quantum yield measured in an integrating hemisphere (QE-2100, Otsuka Electronics). The QD in toluene solution was prepared in an open-type quartz cell, and the absorbance was adjusted to approximately 0.70 ± 0.05 to offset the effect of the QD concentration. The photoluminescence spectrum of the QD solution was collected at room temperature from 500 nm to 800 nm at a xenon-lamp excitation of 150 W (wavelength of 450 nm; wavelength range: -20 nm to $+20 \text{ nm}$). The slit was fixed at 0.6 nm and the data was accumulated 4 times to generate the average value. We measured the quantum yield of the rhodamine 6G, which showed the value of $96 \pm 0.6\%$ in agreement with the literature (quantum yield = 95%). We prepared each QD in different batches according to the same recipe to confirm the reproducibility of the synthetic process.

Time-resolved photoluminescence decay and spectra were obtained with a fluorescence lifetime spectrometer (FluoTime 300, PicoQuant) by means of a time-correlated single-photon-counting data acquisition. To excite the samples, a picosecond pulsed diode laser (LDH-P-FA-530B, PicoQuant) with repetition rate of 800 kHz was used. The photons were collected using a photomultiplier (PMA-C-192, PicoQuant) connected to a time-correlated single-photon-counting board (TimeHarp 260 Pico, PicoQuant) with the overall instrumental response function about 200 ps (FWHM). To reconstruct the time-resolved photoluminescence spectra, probe wavelength-dependent photoluminescence decay profiles were obtained with spectral intervals of 5 nm. All photoluminescence decay profiles were collected with a fixed acquisition time. TEM analysis was performed on a Titan ChemiSTEM electron microscope operated at 200 keV. X-ray photoelectron spectroscopy measurements were carried out using a Quantera II system equipped with a mono-chromatized Al K α source. X-ray diffraction spectra were recorded by a D8 Advance (Bruker) using a Cu K α source. Pyrolysis gas chromatography–mass spectrometry (GC-MS) was performed on a 7890B/5977A (Agilent). Thermogravimetric analysis (TGA) was performed from 20 °C to 600 °C at a heating rate of 10 °C min^{-1} under N_2 using a Trios V3.2 (TA Instruments). Hydrodynamic particle sizes of the QDs in octane were determined using a particle size analyser

(ELSZ-2000, Otsuka). NMR measurements of HA and $\text{Zn}(\text{HA})_2$ were performed with a Bruker Avance III 600.

Transient absorption measurements

The femtosecond transient absorption spectroscopy was employed at the controlled excitation photon fluence to monitor and determine the biexciton Auger recombination lifetimes in the QDs. The fundamental beam (800 nm) pumped by a Ti:sapphire laser system was modified by the optical parametric amplifier to generate the pump beam energy of 620 nm to selectively excite the band-edge states of the QDs and to prevent hot-carrier-related dynamics from complicating the Auger recombination dynamics. Pump fluence was carefully controlled to generate the desired number of excitons, $\langle N_x \rangle$. By probing the 1S excitonic bleach kinetics at different pump fluences, we observed that the amplitude of the fast decaying component (tens of picoseconds) systematically increased as the pump fluence was intensified. The extracted biexciton Auger dynamics in Fig. 3b were obtained by first tail-normalizing the decay dynamics at high and low pump fluences, and then subtracting the decay dynamics of $\langle N_x \rangle = 2$ by the lowest pump fluence decay dynamics ($\langle N_x \rangle < 1$) to isolate the power-dependent decay component. The isolated biexciton components were well fitted by a single-exponential decay function, yielding the time constants of 23 ps, 38 ps and 49 ps for QD-1, QD-2 and QD-3R, respectively.

Single-dot studies and the $g^{(2)}$ statistics

The single-dot study was performed by using the scanning confocal microscopy setup with circularly polarized picosecond pulsed excitation light at 532 nm and a repetition rate of 1 MHz. The laser beam was focused on the focal plane of the high numerical aperture oil immersion objective. Highly diluted QD solutions were spin-cast on a cleaned coverslip, and the photons emitted from individual dots were collected by the same objective and led to the detection part of the setup. The collected photoluminescence was split by a 50:50 non-polarizing cube beam splitter and detected by two avalanche-photodiode-based single photon counting modules in a Hanbury–Brown–Twiss geometry. One of the two signals was delayed by applying an electronic delay of 5 μs in order to detect multiple photons generated by the same excitation pulse. The time intervals between photons were measured by a time-to-amplitude converter, and the second-order intensity correlation ($g^{(2)}$) histogram of inter-photon arrival time was collected with the spacing determined by the laser repetition rate (1 MHz). The measurements were carried out under low excitation intensity ($\langle N_x \rangle \approx 0.01$), under which the $g^{(2)}$ measurements provide an immediate evaluation of the relative biexciton emission efficiency, q_{xx} . The area of central peaks ($g^{(2)}(0)$) accounts for the coincidence events or the biexciton emission, while the lateral peak area ($g^{(2)}(T_{\text{sep}})$) are the single-exciton emission events; where T_{sep} is the pulse-to-pulse separation time. Estimation of relative biexciton emission efficiency q_{xx} was calculated by dividing the central peak area, $g^{(2)}(0)$, by the average lateral peak area, $g^{(2)}(T_{\text{sep}})$.

Determination of QD occupation numbers

In both the TA and single-dot spectroscopic measurements, the pump fluence was controlled to generate the desired average number of excitons per QD, $\langle N_x \rangle$. This value can be quantified by the relation $\langle N_x \rangle = \sigma_{\text{abs},j} \phi$, where σ_{abs} is the absorption cross-section of QDs in cm^2 and j is the pump photon flux (photons per cm^2 pulse). Thus, given the values of absorption cross-sections at the pump wavelength, we could determine the QD occupation numbers at varying pump fluences.

Effective mass approximation modelling

The electron effective masses of InP, ZnSe and ZnS used in the calculation are 0.08, 0.16, and 0.39, respectively. The hole effective masses of InP, ZnSe, and ZnS are 0.6, 0.75 and 1.76 in units of the electron free mass, respectively. The bulk bandgaps of InP, ZnSe and ZnS are 1.42 eV, 2.72 eV and 3.68 eV, respectively. The valence band edge difference

between InP and ZnSe is 0.56 eV, and the difference between ZnSe and ZnS is 0.52 eV.

Fabrication and characterization of QD-LEDs

The glass substrate with a pre-patterned ITO anode (thickness 150 nm) was cleaned with acetone and isopropyl alcohol in an ultrasonic bath and then dried by an isopropyl alcohol dryer. Oxygen plasma treatment was undertaken for 20 min using an ultraviolet ozone cleaner (Jelight, UVO144AX-220). A PEDOT:PSS dispersion was spin-coated onto the ITO pre-patterned substrate with multi-steps of 500 rpm (5 s) and 3,000 rpm (50 s), followed by pre-baking at 110 °C for 10 min in air to remove any residual moisture and hard-baking at 150 °C for 30 min in a nitrogen-filled glovebox (<0.1 ppm H_2O , <0.1 ppm O_2). 0.7 weight per cent of TFB in o-xylene was spin-coated at 2,000 rpm and baked at 150 °C for 30 min in the glove box. The QD solution in octane (15 mg ml^{-1}) was spin-coated at 3,000 rpm, followed by baking at 120 °C for 30 min. ZnMgO nanoparticles were prepared according to literature² and the solution (70 mg ml^{-1} in ethanol) was spin-coated at 4,000 rpm and thermally treated at 140 °C for 30 min under N_2 . Al was deposited by a thermal evaporator with a rate of 1 \AA s^{-1} under a vacuum pressure less than 5×10^{-7} Torr as the cathode, and the device was encapsulated. All solution layers of the electron-only device (ITO (150 nm)/ZnMgO (30 nm)/QD (20 nm)/ZnMgO (30 nm)/Al (100 nm)) and hole-only device (ITO/PEDOT:PSS (35 nm)/TFB (25 nm)/QD (20 nm)/MoO₃ (10 nm)/Al (100 nm)) were processed according to the method described above and the vacuum-deposition layer of the hole-only device was fabricated by thermal evaporation. The current–voltage–luminance characteristics of devices were measured using a system incorporating the spectroradiometer (CS-2000A, Konica Minolta) and a source meter unit (SMU 2635B, Keithley instruments). The unit device was mounted on a 2-wired connection jig installed in a black box with ambient luminance of 0.001 cd m^{-2} or less. The current–voltage sweep was performed by a step voltage of 0.2 V and current density was measured for our pixel areas with a source delay of 0.1 s. The spectroradiometer was tuned to measure within a few seconds at low luminance of 1 cd m^{-2} or less and 1 s at a high luminance of $10,000 \text{ cd m}^{-2}$ or more so that the device did not experience degradation during the current–voltage–luminance sweeps. The aperture size of CS-2000A was set to 0.2° to cover all of the wavelength reference ranges. The EQE was calculated on the assumption that light would have Lambertian distribution.

The lifetime results of the QD-LEDs were obtained using a commercialized multi-channel lifetime test system with an embedded photodiode located in a temperature- and humidity-controlled chamber. The ionization potential was measured by a photoelectron spectro-photometer (AC-3, Riken Keiki) in air.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

- Cao, F. et al. A layer-by-layer growth strategy for large-size InP/ZnSe/ZnS core-shell quantum dots enabling high-efficiency light-emitting diodes. *Chem. Mater.* **30**, 8002–8007 (2018).
- Jo, J.-H. et al. High-efficiency red electroluminescent device based on multishelled InP quantum dots. *Opt. Lett.* **41**, 3984–3987 (2016).
- Wang, H. C. et al. Cadmium-free InP/ZnSeS/ZnS heterostructure-based quantum dot light-emitting diodes with a ZnMgO electron transport layer and a brightness of over 10000 cd m^{-2} . *Small* **13**, 1603962 (2017).
- Lim, J. et al. Highly efficient cadmium-free quantum dot light-emitting diodes enabled by the direct formation of excitons within InP@ZnSeS quantum dots. *ACS Nano* **7**, 9019–9026 (2013).
- Shen, H. et al. Visible quantum dot light-emitting diodes with simultaneous high brightness and efficiency. *Nat. Photon.* **13**, 192–197 (2019).
- Kim, Y. et al. Bright and uniform green light emitting InP/ZnSe/ZnS quantum dots for wide color gamut displays. *ACS Appl. Nano Mater.* **2**, 1496–1504 (2019).
- Park, Y., Lim, J., Makarov, N. & Klimov, V. I. Effect of interfacial alloying versus “volume scaling” on Auger recombination in compositionally graded semiconductor quantum dots. *Nano Lett.* **17**, 5607–5613 (2017).

Acknowledgements We thank B. G. Chae, H. Park and H. Heo for their assistance with TEM, GC-MS and ICP-AES.

Author contributions The synthesis of the QDs and structural analysis were performed by Y.-H.W. and H.J. The QD-LEDs were fabricated and characterized by O.C. and D.-Y.J. The HR-STEM images were obtained by J.L. The photophysical properties of the QDs were studied by T.K., T.K., H.C. and D.K. This research was designed and coordinated by E.J. This manuscript was written by Y.-H.W. and E.J. in consultation with all authors.

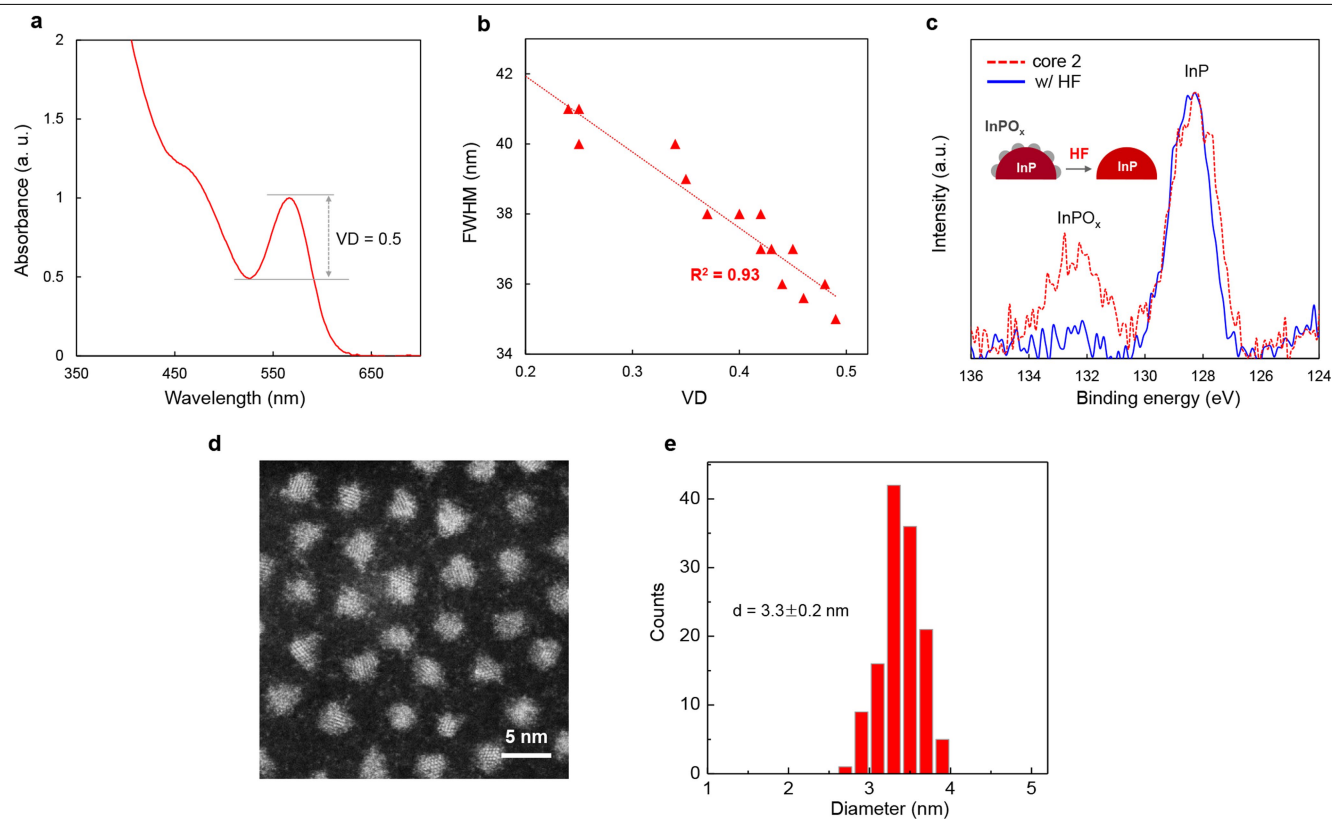
Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.J.

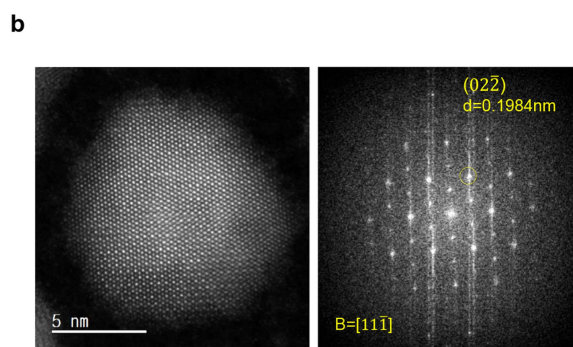
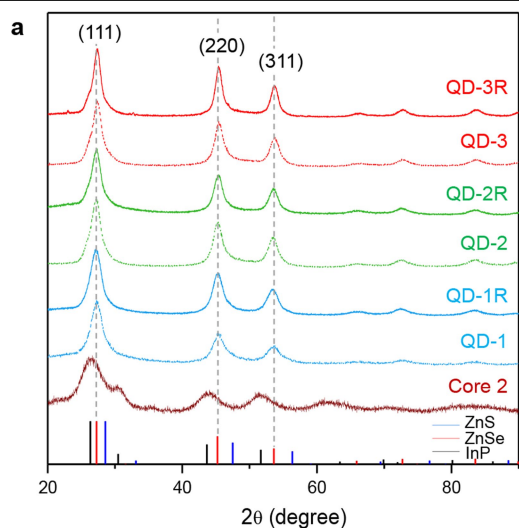
Peer review information *Nature* thanks Peter Reiss and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



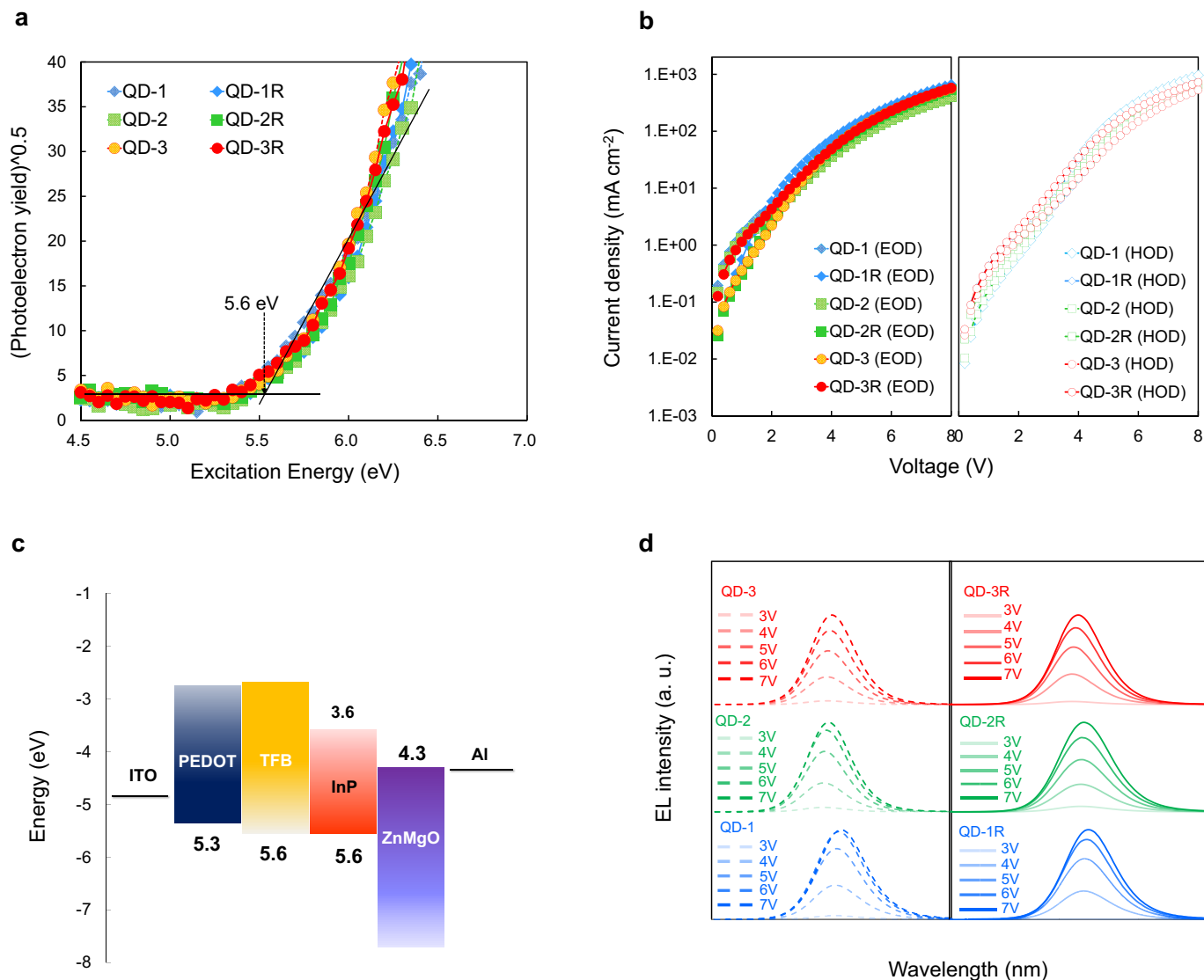
Extended Data Fig. 1 | Characteristics of the InP core QD. **a**, Ultraviolet-visible absorption spectrum of InP core 2 and a valley depth (VD) defined as $VD = 1 - (Abs_{min}/Abs_{max})$. **b**, Plot of VD of InP core versus FWHM of InP/ZnSe/ZnS QDs prepared with each respective InP core. R^2 is the proportion of the variance

from the regression model. **c**, X-ray photoelectron spectroscopy profiles of P_{2p} of InP core 2 before and after the HF treatment (inset, schematic structure of the InP surface before and after HF addition). **d**, High-resolution-STEM image. **e**, Histogram of the diameters of core 2.



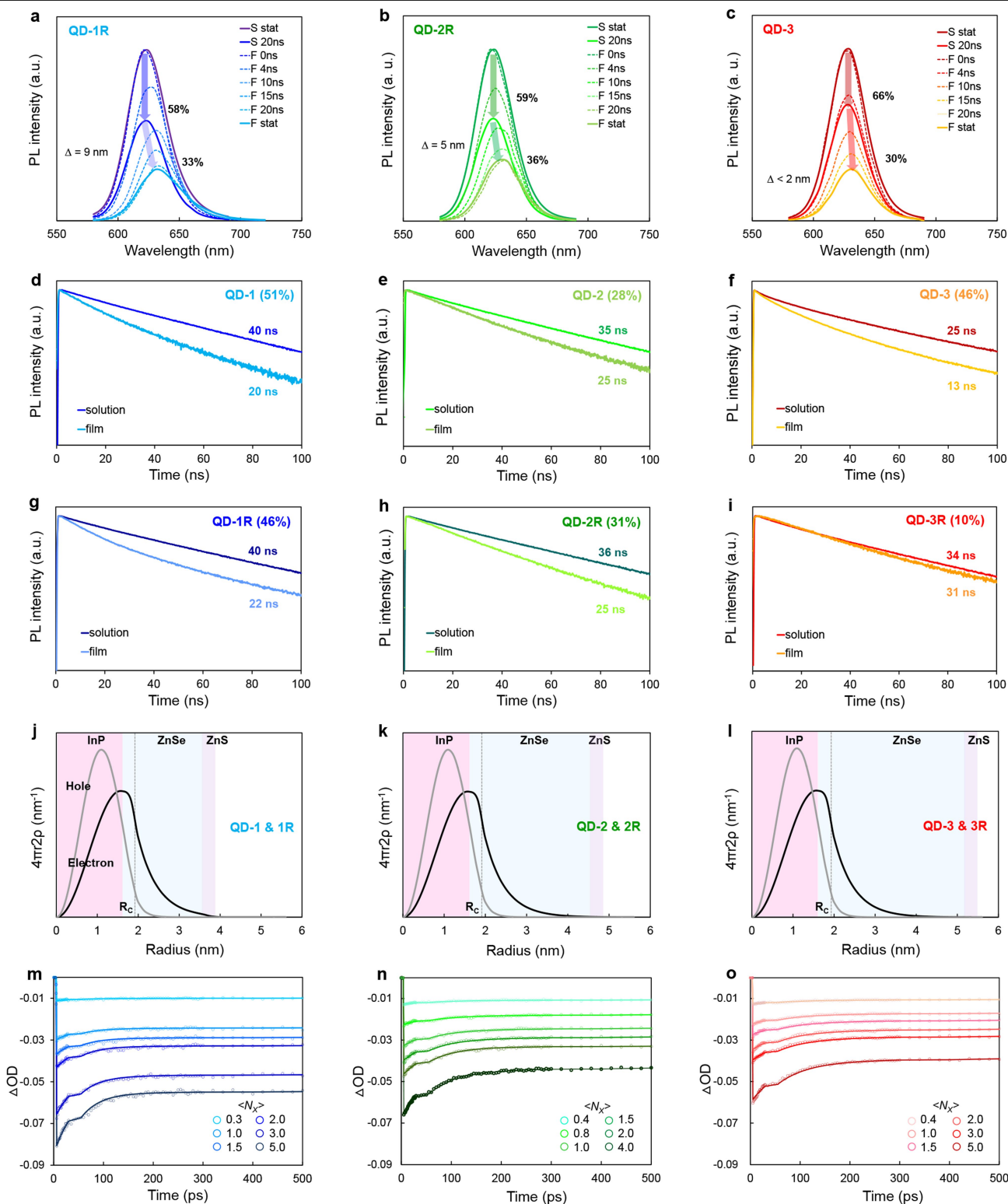
Extended Data Fig. 2 | Crystalline structures of the QDs. a, X-ray diffraction patterns of the core 2, QD-1, QD-1R, QD-2, QD-2R, QD-3 and QD-3R. **b,** High-resolution-STEM image and corresponding fast Fourier transform pattern on

the $[11\bar{1}]$ zone axis (B) of QD-3R. The atomic distance d of the crystalline facet of (02 $\bar{2}$) was 0.1984 nm.



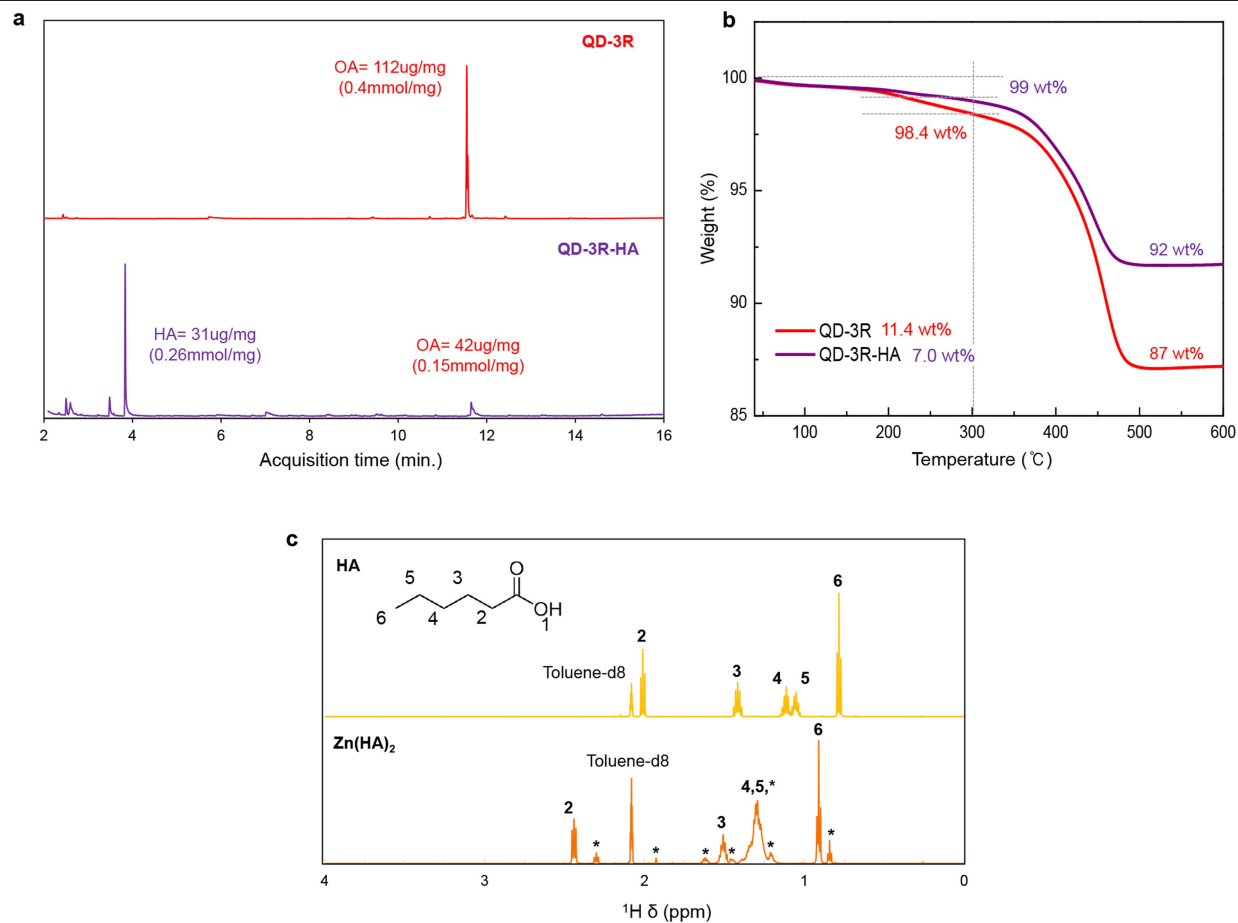
Extended Data Fig. 3 | Electronic structures of the InP/ZnSe/ZnS QDs and their performances in the devices. a, Valence band maximum measurement with photoelectron microscopy of QD-1, 2, 3, 1R, 2R and 3R. **b,** Current density–voltage profiles of the hole-only devices and electron-only devices using QD-1,

2, 3, 1R, 2R and 3R. **c,** Energy level diagram of the reference LED device. **d,** Electroluminescence spectra of the QD-LEDs using QD-1, 2, 3, 1R, 2R and 3R, depending on the operation voltages.



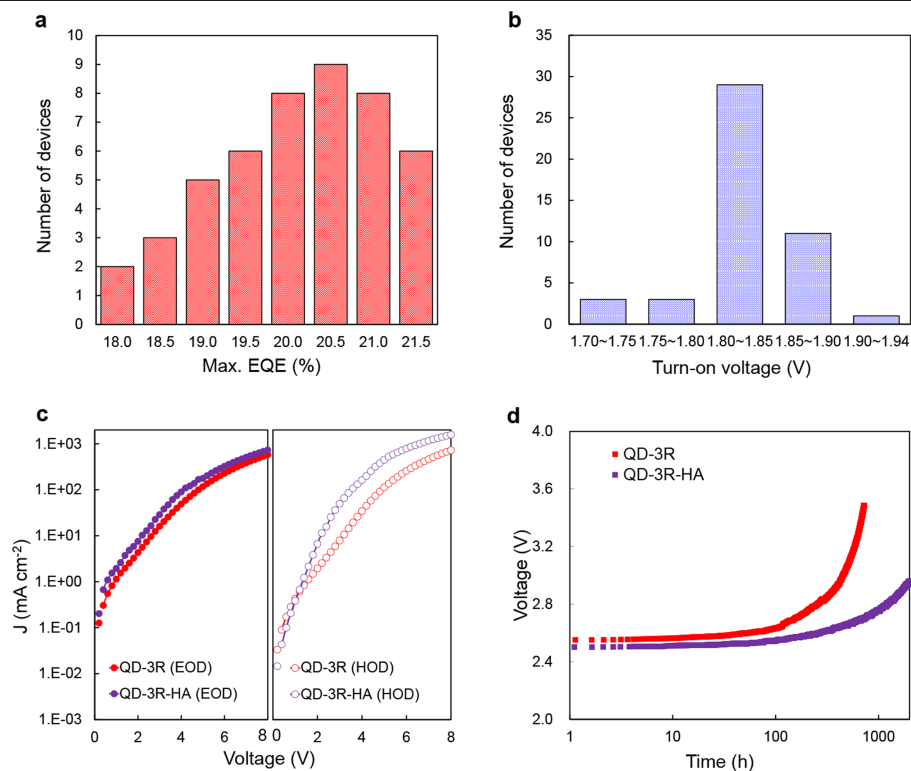
Extended Data Fig. 4 | Photophysical characteristics based on the time-resolved spectroscopy and electron-hole distributions. **a–c**, Static photoluminescence spectra of the QD solution (S) and the QD film (F, solid line) together with the dynamic photoluminescence spectra (dashed lines) of the QD film collected at each decay time (0–20 ns) of: **a**, QD-1R; **b**, QD-2R; and **c**, QD-3. **d–i**, Spectrally integrated photoluminescence decay profiles of the QD in solution and film of: **d**, QD-1; **e**, QD-2; **f**, QD-3; **g**, QD-1R; **h**, QD-2R; and **i**, QD-3R. The number next to the label of each photoluminescence decay profile

indicates the average photoluminescence lifetime. From the average photoluminescence lifetimes, the energy transfer efficiency was calculated as $E_t = 1 - \tau_{film}/\tau_{sol}$ and is given next to the sample name. **j–l**, The electron and hole distributions were calculated with an effective mass approximation of: **j**, QD-1, 1R; **k**, QD-2, 2R; and **l**, QD-3, 3R. (R_c is the effective core radius, defined as a radius of the sphere that confines each electron or hole carrier with 90% probability based on the effective mass approximation calculation.) **m–o**, Transient absorption dynamics of **m**, QD-1, **n**, QD-2, and **o**, QD-3R.



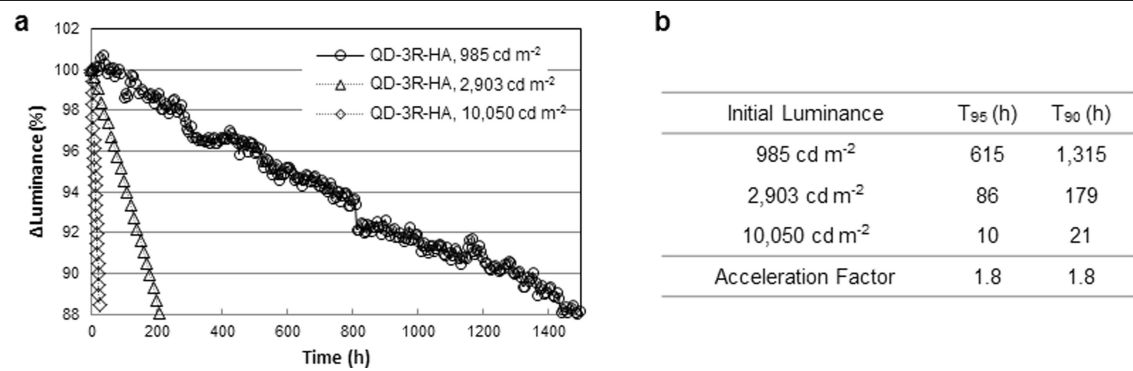
Extended Data Fig. 5 | Ligand characterization before and after the exchange reaction. a, b, GC-MS data (a) and TGA data (b) of QD-3R and QD-3R-HA before and after the ligand exchange with quantitative analytical data. **c,** ^1H

NMR spectra of HA and $\text{Zn}(\text{HA})_2$ in toluene- d_8 . $\text{Zn}(\text{HA})_2$ is dissolved in TOA and the asterisk indicates TOA solvent peaks.



Extended Data Fig. 6 | Performance of the QD-LEDs. **a, b**, Histograms of maximum EQEs (**a**) and turn-on voltages at 0.1 cd m^{-2} (**b**) measured from 47 QD-LEDs with QD-3R-HA. **c**, Current density-voltage profiles of electron- and hole-

only devices fabricated with QD-3R and QD-3R-HA. **d**, Voltage versus operation time for the QD-LEDs with QD-3R and QD-3R-HA operated at a constant current fixed initially at about $1,000 \text{ cd m}^{-2}$.



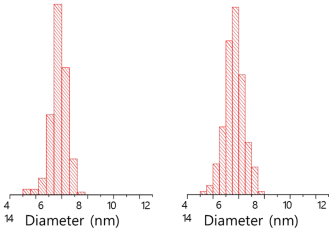
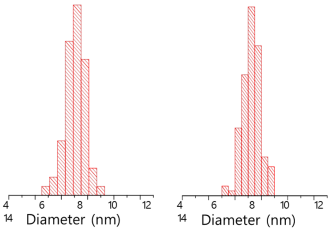
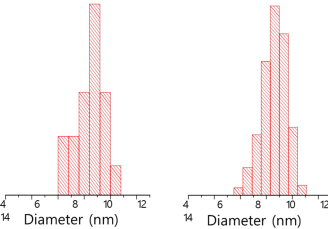
Extended Data Fig. 7 | Lifetime measurements and acceleration factor. a, Lifetimes of the QD-LED with QD-3R-HA at 985 cd m^{-2} , 2,903 cd m^{-2} and 10,050 cd m^{-2} . **b,** Acceleration factor calculated for different conditions by fitting T_{95} and T_{90} data.

Extended Data Table 1 | Performances of the QD-LEDs with previously reported InP-based QDs and CdSe-based QDs and those developed in this work

QDs	Colour	PL peak (nm)	PL FWHM (nm)	QY (%)	Diameter (nm)	Max. EQE (%)	Max. Luminance (cd m ⁻²)	Lifetime	Reference
InP	Red	630	35	100	10.8	21.4	100,000	T50 = 1,000,000 h @100 cd m ⁻²	This work
		618	42	93	7.8	12.2	10,000	-	J. Am. Chem. Soc. 2019 ¹⁰
		607	48	73		6.6	1,700	-	Chem. Mater. 2018 ²⁵
	Green	607	63	82	7.2	2.5	2,849	-	Opt. Lett. 2016 ²⁶
		525	65	70	7.4		10,490	-	Small 2017 ²⁷
		505	50	72	4.5	3.46	3,900	-	ACS Nano 2013 ²⁸
CdSe	Red	600		90	11	21.6	356,000	T50 = 1,600,000 h @100 cd m ⁻²	Nat. Photon. 2019 ²⁹
		627	21	84	9.8	15.1	< 30,000	T95 = 2,320 h @1,000 cd m ⁻²	Nature Commun. 2018 ⁹
					8	12	21,000	T50=300,000 h @100 cd m ⁻²	Nat. Photon. 2015 ⁸
		623	30	80	8.3	7.4	105,870	-	Adv. Mater. 2014 ¹⁸
				> 90		20.5	42,000	T50 = 100,000 h @100 cd m ⁻²	Nature, 2014 ⁴
	Green	615			6	18	50,000	-	Nat. Photonic 2013 ³
		525		90	6.7	22.9	614,000	T50 = 1,760,000 h @100 cd m ⁻²	Nat. Photon. 2019 ²⁹
						21	40,000	T50 = 280,000 h @100 cd m ⁻²	J. SID 2015 ⁵
	Blue	533		75	7~8	14.5		T50 = 90,000 h @100 cd m ⁻²	Nat. Photon. 2015 ⁸
		475		73		8.05	62,600	T50 = 7,000 h @100 cd m ⁻²	Nat. Photon. 2019 ²⁹
	Blue	466		87	12~13	19.8	4,890	T50 = 47.4 h @100 cd m ⁻²	ACS AMI 2017 ⁶
					8	10.7	4,000	T50 < 1,000 h @100 cd m ⁻²	Nat. Photon. 2015 ⁸

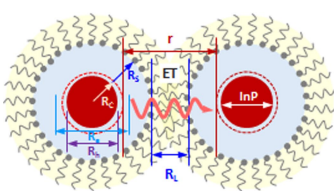
Data are taken from refs. ^{3-6,8-10,18,25-29}, PL, photoluminescence. QY, quantum yield.

Extended Data Table 2 | Characteristics of the InP/ZnSe/ZnS QDs with different shell structures

		QD-1	QD-1R	QD-2	QD-2R	QD-3	QD-3R
PL peak (FWHM)		631 (35)	629 (39)	627 (35)	628 (37)	630 (35)	631 (35)
QY (%)		98 ± 1.3	100 ± 0.9	92 ± 2.5	100 ± 1.7	75 ± 1.9	100 ± 1.4
TEM Size (nm)		7.4 ± 0.8	7.7 ± 0.7	9.2 ± 1.1	9.2 ± 0.8	10.8 ± 1.1	10.8 ± 0.9
Histogram							
ICP (Mole ratio)	In/In	1.0	1.0	1.0	1.0	1.0	1.0
	P/In	0.7	0.8	0.8	0.9	0.9	0.9
	Zn/In	12.5	10.8	24.5	20.8	37.0	36.5
	Se/In	7.9	8.3	17.5	16.5	27.1	28.4
	S/In	2.8	1.5	3.8	2.2	5.2	4.7
Calculated based on the ICP	ZnSe thickness (nm)	1.9	1.9	2.8	2.7	3.5	3.6
	ZnS thickness (nm)	0.3	0.2	0.3	0.2	0.3	0.2
	diameter (nm)	7.7	7.5	9.5	9.1	10.9	10.9
Circularity		0.74	0.81	0.70	0.84	0.62	0.84
Solidity		0.89	0.95	0.89	0.96	0.87	0.96

Optical properties, particle sizes obtained from TEM images (histogram), elemental compositions obtained from ICP analysis, shell thickness, and total diameter calculated based on the ICP using the method of ref. ³⁰, circularity (degree of similarity to a perfect circle = 4π(area)/(perimeter)²) and solidity (degree of convexity = area/convex area) calculated from the TEM images.

Extended Data Table 3 | FRET efficiency for the InP/ZnSe/ZnS QDs

Scheme	Parameter	QD-1	QD-2	QD-3R
	Orientation factor, κ	0.816	←	←
	Solution QY, Φ	0.98	0.92	1.00
	Spectral overlap integral, $J(\lambda)$	9.5.E+15	1.1.E+16	1.2.E+16
	Refractive index, n	2.22	2.2	2.2
	Critical distance, R_0 [nm]	5.3	5.4	5.6
	Effective electron radius, R_e [nm]	2.30	2.32	2.32
	Effective hole radius, R_h [nm]	1.63	1.63	1.63
	Effective core radius, R_c [nm]	1.93	1.94	1.94
	Effective shell thickness, R_s [nm]	1.77	2.66	3.46
	Spacing between QDs by ligands, R_L [nm]	1.92	←	←
	Distance between QDs, r [nm]	5.46	7.24	8.84
	FRET efficiency, E_f [%]	46.2	15.4	5.9
	Film QY [%]	69	84	93

R_0 is calculated as: $R_0 = 0.211(\kappa^2 \Phi J(\lambda) n^{-4})^{1/6}$ (ref. ²⁰); where n is a volume-weighted average of the refractive indices of InP, ZnSe, ZnS, and oleic acid²¹. E_f is calculated as $E_f = R_0^6 / (R_0^6 + r^6)$ (ref. ²⁰), where $r = 2R_s + R_L$; R_c is calculated using $V_{eff} = 8\pi/3 \times R_e^2 \times R_h / (R_e + R_h) = 4\pi/3 R_c^3$, where V_{eff} is the effective excitonic volume²¹. R_s = QD diameter minus R_c , and R_L is directly measured from the high-resolution SEM image of the QD films. All the distances used in the table are described in the schematic image.

Domino electroreduction of CO₂ to methanol on a molecular catalyst

<https://doi.org/10.1038/s41586-019-1760-8>

Yueshen Wu^{1,2}, Zhan Jiang³, Xu Lu^{1,2}, Yongye Liang^{3*} & Hailiang Wang^{1,2*}

Received: 25 November 2018

Accepted: 2 October 2019

Published online: 27 November 2019

Electrochemical carbon dioxide (CO₂) reduction can in principle convert carbon emissions to fuels and value-added chemicals, such as hydrocarbons and alcohols, using renewable energy, but the efficiency of the process is limited by its sluggish kinetics^{1,2}. Molecular catalysts have well defined active sites and accurately tailorable structures that allow mechanism-based performance optimization, and transition-metal complexes have been extensively explored in this regard. However, these catalysts generally lack the ability to promote CO₂ reduction beyond the two-electron process to generate more valuable products^{1,3}. Here we show that when immobilized on carbon nanotubes, cobalt phthalocyanine—used previously to reduce CO₂ to primarily CO—catalyses the six-electron reduction of CO₂ to methanol with appreciable activity and selectivity. We find that the conversion, which proceeds via a distinct domino process with CO as an intermediate, generates methanol with a Faradaic efficiency higher than 40 per cent and a partial current density greater than 10 milliamperes per square centimetre at −0.94 volts with respect to the reversible hydrogen electrode in a near-neutral electrolyte. The catalytic activity decreases over time owing to the detrimental reduction of the phthalocyanine ligand, which can be suppressed by appending electron-donating amino substituents to the phthalocyanine ring. The improved molecule-based electrocatalyst converts CO₂ to methanol with considerable activity and selectivity and with stable performance over at least 12 hours.

On the basis of the Sabatier principle, the binding energy of CO, $E_B(\text{CO})$, is often used as a descriptor to understand the different catalytic selectivities of metal surfaces in the electroreduction of CO₂^{4–7}. On metals that bind CO too weakly (for example, Ag and Au, which have $E_B(\text{CO})$ values that are relatively positive), CO easily desorbs upon formation and is thus the major product of CO₂ reduction (Fig. 1a). For metals that bind CO too strongly (for example, Ni and Pt), a very negative $E_B(\text{CO})$ makes further reduction of adsorbed CO (CO*) only possible at very negative potentials, where the competing H₂ evolution reaction dominates. As a result, H₂ is the major reduction product on these metal surfaces⁴ (Fig. 1a). To enable deeper CO₂ reduction to hydrocarbons or oxygenates, a moderate $E_B(\text{CO})$ is required, so that CO* stays bound to the catalytic site and its reduction can proceed with a reasonably low energy barrier⁶. In fact, Cu is currently the only metal that can catalyse CO₂ electroreduction to more deeply reduced products with appreciable selectivity^{6,7}. In search for an electrocatalyst other than metals that can reduce CO₂ by more than two electrons, we consider that there may be a suitable candidate molecule that (1) is capable of catalysing CO₂-to-CO conversion and (2) has a moderate binding strength for CO. Recent computational studies of M–N₄ molecular structures (a metal centre coordinated with four nitrogen atoms), which are active in catalysing the CO₂ electroreduction to CO^{8–11}, have shown that $E_B(\text{CO})$ on these sites can vary substantially with the identity of the metal ion^{12–14}.

For example, CO binding is strong on Fe–N₄, moderate on Co–N₄ and weak on Ni–N₄, spanning an energy range of about 1.2 eV (Fig. 1a, Extended Data Table 1). Interestingly, $E_B(\text{CO})$ for Co–N₄ is similar to that of Cu. Although such $E_B(\text{CO})$ values may not be directly put into the context of the scaling relations and reactivity trends established for metal surfaces, because the catalysts are molecular in nature and their $E_B(\text{CO})$ is influenced to some extent by the peripheral structure and oxidation state of the metal centre^{15,16}, this $E_B(\text{CO})$ trend still shows a dependence on the identity of the metal centre and is well correlated with recently published experimental results^{17,18}. This points to the exciting possibility that CO₂ may be deeply reduced beyond CO on M–N₄-based electrocatalyst materials.

To explore this possibility, we chose iron phthalocyanine (FePc), cobalt phthalocyanine (CoPc) and nickel phthalocyanine (NiPc) molecules supported on carbon nanotubes (CNTs) as catalysts for the initial screening. Our noncovalent anchoring strategy⁸ enables catalytic molecules to be highly dispersed on the surface of a highly conductive network and renders heterogenized molecular catalysts that may be able to overcome some of the limitations of homogeneous electrocatalysts: in homogeneous electrochemical CO₂ reduction, a catalyst molecule diffuses to the electrode to accept one or two electrons and then diffuses away from the surface to react with CO₂ in the solution^{1,3}, which makes it difficult to transfer multiple electrons to a CO₂ molecule to form deeply reduced products.

¹Department of Chemistry, Yale University, New Haven, CT, USA. ²Energy Sciences Institute, Yale University, West Haven, CT, USA. ³Department of Materials Science and Engineering, Guangdong Provincial Key Laboratory of Energy Materials for Electric Power, Southern University of Science and Technology, Shenzhen, China. *e-mail: liangyy@sustech.edu.cn; hailiang.wang@yale.edu

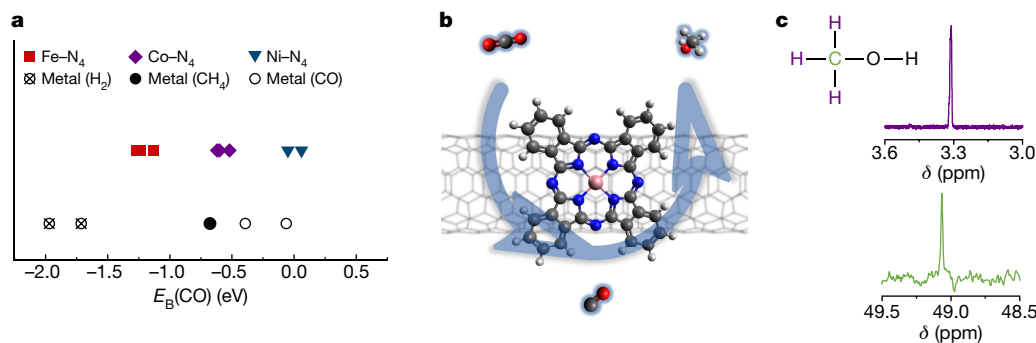


Fig. 1 | A domino electrocatalytic process of CO₂-to-MeOH conversion on CoPc/CNT, discovered through a catalyst search guided by the Sabatier principle. a, Computed CO binding energies on metal surfaces and on M-N₄ moieties (data from refs. ^{6,7,13,14,24}). Metals are classified according to their major product (denoted inside parentheses) for the electroreduction of CO₂.

b, Domino process of CO₂-to-MeOH conversion via CO, catalysed by CoPc supported on carbon nanotubes (CNT). Colour code: hydrogen, light grey; carbon, dark grey; nitrogen, blue; oxygen, red; cobalt, pink. **c**, Zoomed-in regions of sample ¹H NMR and ¹³C NMR spectra of the electrolyte that confirm MeOH production.

Measured in 0.1 M KHCO₃ aqueous electrolyte, all three MPC/CNT (M = Fe, Co or Ni) catalysts show a decent selectivity for CO generation in the medium overpotential range: CoPc/CNT and NiPc/CNT both achieve a maximum Faradaic efficiency (FE) of about 95% for CO production, whereas FePc/CNT exhibits a lower CO selectivity of about 80% (Extended Data Fig. 1, Fig. 2a). At more negative electrode potentials, the CO production process on both NiPc/CNT and FePc/CNT is taken over by H₂ evolution. No other gaseous products are detected (Extended Data Fig. 2a, b). Interestingly, CoPc/CNT behaves differently: it generates methanol (MeOH; Fig. 2a), which is confirmed to be the only liquid-phase product by both ¹H and ¹³C nuclear magnetic resonance (NMR) spectroscopy (Fig. 1c, Extended Data Fig. 2c, d). MeOH production on CoPc/CNT onsets at about -0.82 V with respect to the reversible hydrogen electrode (RHE) and reaches its highest FE of 44% (see Extended Data Fig. 2e for the detailed quantification procedure) and largest partial current density of 10.6 mA cm⁻² (corresponding to a turnover frequency of 1.05 s⁻¹ if all the supported CoPc molecules are counted as active sites) at -0.94 V (Fig. 2a, b), as measured with a 1-h electrolysis. To the best of our knowledge, this is the first example of a transition-metal-based molecular electrocatalyst producing MeOH from CO₂ with an appreciable yield (Extended Data Table 2). A control experiment using N₂ as the feed gas produces only H₂ and a trace amount of CO (Extended Data Fig. 3), thereby confirming that the MeOH indeed comes from reduction of CO₂.

One key feature of our catalyst is the molecular-level dispersion of CoPc on CNTs, as evidenced by scanning transmission electron microscopy (STEM) images recorded with a high-angle annular dark-field (HAADF) detector and the corresponding elemental mapping obtained with energy dispersive spectroscopy (EDS). N was found to be uniformly distributed on the CNTs (Fig. 2d, e). The atomic-resolution Z-contrast image reveals the distribution of Co atoms on a CNT and directly confirms the molecular-level dispersion of CoPc (Fig. 2f). Such a high level of dispersion was found to be necessary for the selective CO₂-to-MeOH conversion. When CoPc without CNT supports was directly deposited from its solution onto a carbon fibre paper, the resulting CoPc electrode showed a much lower current density than CoPc/CNT at each measured potential and produced H₂ and CO as the only products (Extended Data Fig. 4a, b). When CoPc was physically mixed with CNTs, both the selectivity and the activity of MeOH production were considerably lower than in the case of the CoPc/CNT hybrid (Extended Data Fig. 4d), even though the content of CoPc in the mixture was ten times higher than that in the hybrid. The MeOH production rate can be inversely correlated with the charge-transfer resistance derived from electrochemical impedance spectroscopy measurements under the working conditions (Extended Data Fig. 4c). A physical mixture of CoPc and another carbon support (Vulcan XC72 or Ketjenblack) also shows much lower FE_{MeOH} and partial current density compared to

the CoPc/CNT hybrid catalyst at -0.94 V versus RHE (Extended Data Fig. 4d). Taken together, these results support that effectively dispersing CoPc molecules on highly conductive supports can expose their catalytic reactivity and enable more active and selective CO₂-to-MeOH conversion than CoPc aggregates.

Because MeOH emerges as a product at the expense of CO as the electrode potential is polarized more negatively than -0.77 V, we hypothesize that CO is an intermediate in the CO₂-to-MeOH process. To test this hypothesis, we carried out electroreduction of CO with the CoPc/CNT electrode in the same electrolyte. MeOH could be detected at potentials more negative than -0.77 V, and FE reached 28% at -0.83 V (Fig. 2c). The activity for the electroreduction of CO therefore indicates that the catalytic CO₂ reduction to MeOH on CoPc/CNT follows a domino process in which CO₂ first undergoes a two-electron reduction to CO, which continues to be reduced to MeOH through a four-electron–four-proton process (Fig. 1b). The fact that the onset potential for MeOH formation in the CO electroreduction roughly coincides with that in the CO₂ electroreduction implies that these two reactions share the same potential-limiting step. It is worth emphasizing that CoPc is currently the only catalyst other than Cu that can electrochemically reduce CO with an appreciable current density¹⁹.

The long-term electrolysis results, however, show that the electrocatalytic CO₂ reduction to MeOH on CoPc/CNT is unstable. The average FE_{MeOH} is 44% for the first 1 h and then drops to 26% over the next 4 h. After 5 h of electrolysis, FE_{MeOH} further decreases to a negligible 0.6% (Fig. 3a) while the FE for H₂ evolution increases to -80% (Extended Data Fig. 5a). This deactivation phenomenon may be another contributing factor to the scarcity of reports on CO₂-to-MeOH conversion catalysed by CoPc. Scanning electron microscopy (SEM) images of the electrode after a 12-h electrolysis excluded reductive demetallation of CoPc and formation of Co or CoO_x nanoparticles (Extended Data Fig. 5b, c). The deactivation of the catalyst was then attributed to changes within the molecular structure. Some related porphyrin complexes have been shown to undergo hydrogenation on their pyrrole rings under a protic and reductive environment and manifest a substantial change in their ultraviolet–visible (UV-Vis) absorption profile^{20,21}. We therefore speculate that the deactivation is caused by undesirable reduction of the Pc ligand. To study the degraded catalyst, the CoPc/CNT physical-mixture electrode was used because it has a CoPc content ten times higher than that of the hybrid. After a 2-h electrolysis at -0.94 V, the used CoPc molecules were dissolved by deoxygenated N,N'-dimethylformamide (DMF) and then subjected to spectroscopic characterization. Compared to the UV-Vis spectrum of a fresh CoPc/DMF solution, the used CoPc solution exhibits three new absorption peaks at 420, 460 and 700 nm (Fig. 3b). These new peaks do not pertain to the free-base phthalocyanine (H₂Pc; Extended Data Fig. 6), corroborating the absence

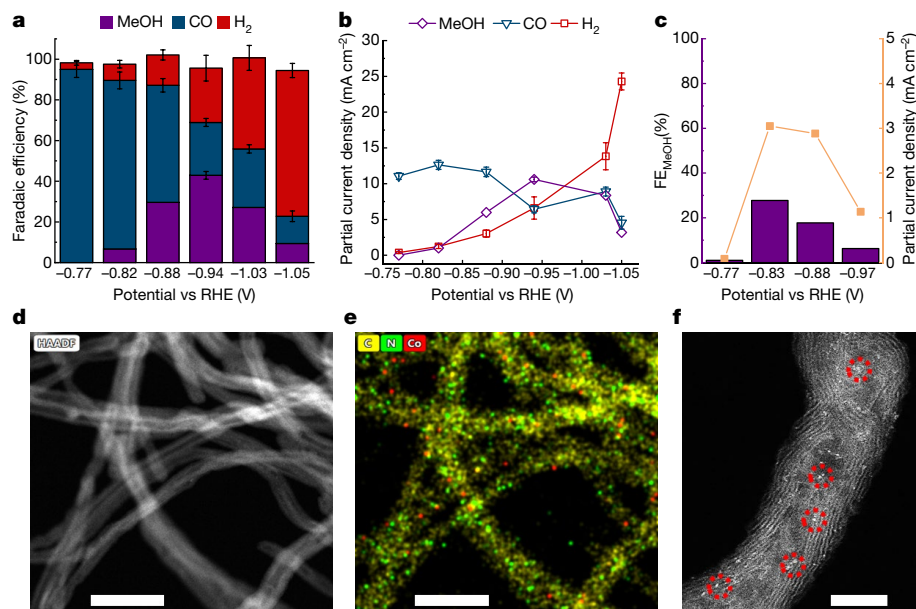


Fig. 2 | Catalytic performance of CoPc molecules supported on CNTs for CO₂ and CO reduction. **a**, Product selectivity (FE; **a**) and partial current densities (**b**) for different products versus electrode potential. Error bars represent one standard deviation from three measurements. **c**, Potential-dependent FE and partial current density for MeOH production from electroreduction of CO

catalysed by CoPc/CNT. **d**, **e**, STEM-HAADF image of CoPc/CNT (**d**) and corresponding overlaid EDS maps of Co, C and N (**e**). **f**, Atomic-resolution STEM-HAADF image of CoPc/CNT. The circled bright spots correspond to the Co centre of individual CoPc molecules. Scale bars, 50 nm (**d**, **e**) and 5 nm (**f**).

of demetallation of the complex. The same absorption peaks appear in the spectrum of CoPc molecules reacted with NaBH₄ in deoxygenated DMF. These emerging absorption features have been shown to be characteristic of singly reduced CoPc, and continuous reduction can eventually lead to hydrogenation of the Pc macrocycle²². These results support our hypothesis that the deteriorating selectivity of the CoPc/CNT catalyst for MeOH production is caused by the detrimental reduction of the ligand.

One major advantage of our heterogenized molecular catalyst system is that its structure can be tailored on the molecular level to improve its catalytic properties. To tackle the reduction-induced deactivation of CoPc/CNT, we introduced four amino groups (-NH₂) to the β positions of the Pc ligand (Fig. 4a). The electron-donating substituents successfully lowered the reduction potential of CoPc (Extended Data Fig. 7). We note that CoPc-NH₂ has a larger conjugation system than aniline, which means that the pK_a of its conjugated acid is probably lower than that of aniline (4.6). Therefore, protonation of the appended amino groups in the CO₂-saturated electrolyte (pH = 6.8) is expected to be minimal. The CoPc-NH₂/CNT catalyst exhibits a similar potential-dependent

behaviour to that of CoPc/CNT. At -1.00 V versus RHE, the conversion of CO₂ to MeOH proceeds with an FE of 32% and an average partial current density of 10.2 mA cm⁻² (Fig. 4b, c), as measured with a 1-h electrolysis. Remarkably, CoPc-NH₂/CNT shows much improved catalytic durability than its unsubstituted counterpart. The measured FE_{MeOH} for a 12-h electrolysis is 28%, comparable to the FE_{MeOH} of the 1-h electrolysis, and the total current density stays between 30 and 33 mA cm⁻² through the entire period (Fig. 4d). After a 2-h electrolysis at -1.00 V, CoPc-NH₂ molecules generate the same UV-Vis spectrum as that of pristine CoPc-NH₂ (Fig. 4e), suggesting that the catalytic structure for MeOH production remains intact under the reaction conditions. The increased durability provided by amino substitution shows the power of ligand engineering in improving the catalytic performance of our heterogenized molecular systems. Notably, the CoPc-NH₂/CNT hybrid material shows a higher selectivity (maximum FE_{MeOH} = 41%) for the electroreduction of CO than for that of CO₂ (Extended Data Figs. 2f, 8).

Several factors contribute to the efficiency of our CoPc/CNT in catalysing CO₂ electroreduction to MeOH. First, CoPc is dispersed as individual molecules on highly conductive CNTs, which is critical to fast and continuous electron delivery to the active site for multi-electron reduction of CO₂. By contrast, a simple mixture of CoPc and CNTs inevitably contains CoPc aggregates, which make it much less efficient for catalysis (Extended Data Fig. 4d). Second, the type of carbon support is important (Extended Data Fig. 4d) because the CoPc molecules are anchored on and accept electrons from the support. Third, CoPc undergoes deactivation at the reductive conditions (Fig. 3a, Extended Data Fig. 5a); it is therefore important to stabilize the active site—for instance, by modifying the Pc ligand—to ensure efficacy during longer-term electrolysis. These factors are additive and together could explain why CO₂-to-MeOH conversion was either not seen at all in other studies using CoPc and its derivatives^{23,24} or was reported to occur with very low activity and selectivity (partial current density <0.05 mA cm⁻², FE <5%)^{25,26}. Lastly, we emphasize that its molecular structure is critical and contains the active site even if support effects are important: other cobalt macrocycle complexes, such as cobalt chlorin and cobalt porphyrin^{27,28}, have not been reported to show MeOH selectivity, even when they are composited with CNTs.

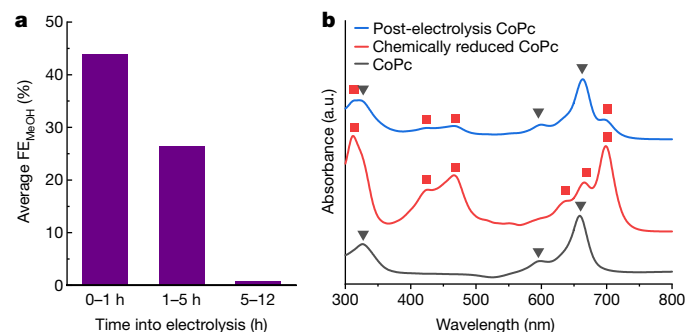


Fig. 3 | Deactivation of CoPc/CNT in long-term electrolysis. **a**, Average FE_{MeOH} of CoPc/CNT for different time intervals in a 12-h electrolysis. **b**, UV-Vis absorption profiles of CoPc, post-electrolysis CoPc and chemically reduced CoPc. Absorption peaks associated with CoPc and reduced CoPc are indicated by grey triangles and red squares, respectively. (a.u., arbitrary units.)

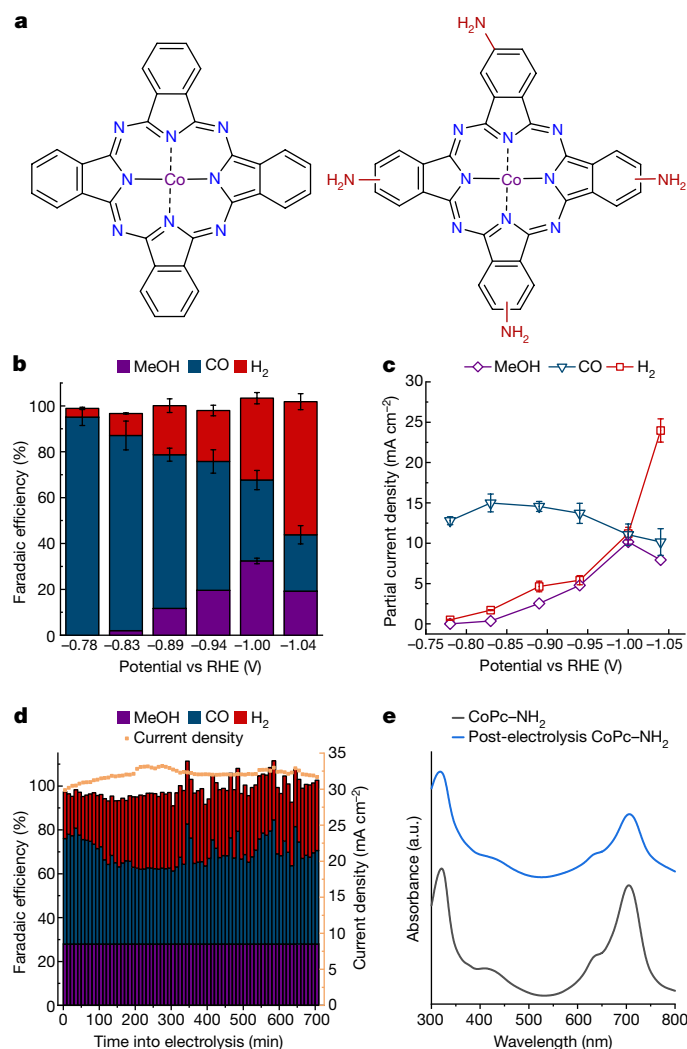


Fig. 4 | Electrochemical performance of CoPc-NH₂/CNT for CO₂ reduction to methanol. a, Structural comparison between CoPc and CoPc-NH₂. **b, c**, Potential-dependent product selectivity (FE; **b**) and partial current density (**c**) for CO₂ electroreduction catalysed by CoPc-NH₂/CNT. Error bars represent one standard deviation from three measurements. **d**, Product selectivity (FE) and total current density for a 12-h electrolysis of CO₂ reduction catalysed by CoPc-NH₂/CNT at -1.00 V versus RHE; the FE_{MeOH} value measured after the electrolysis is shown in striped violet. **e**, Comparison between UV-Vis absorption profiles of CoPc-NH₂ and post-electrolysis CoPc-NH₂.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1760-8>.

- Costentin, C., Robert, M. & Savéant, J.-M. Catalysis of the electrochemical reduction of carbon dioxide. *Chem. Soc. Rev.* **42**, 2423–2436 (2013).
- Schiffer, Z. J. & Manthiram, K. Electrification and decarbonization of the chemical industry. *Joule* **1**, 10–14 (2017).
- Francke, R., Schille, B. & Roemelt, M. Homogeneously catalyzed electroreduction of carbon dioxide—methods, mechanisms, and catalysts. *Chem. Rev.* **118**, 4631–4701 (2018).
- Hori, Y., Wakebe, H., Tsukamoto, T. & Koga, O. Electrocatalytic process of CO selectivity in electrochemical reduction of CO₂ at metal electrodes in aqueous media. *Electrochim. Acta* **39**, 1833–1839 (1994).
- Hori, Y. in *Modern Aspects of Electrochemistry* Vol. 42 (eds Vayenas, C. G. et al.) 89–189 (Springer, 2008).
- Peterson, A. A., Abild-Pedersen, F., Studt, F., Rossmeisl, J. & Nørskov, J. K. How copper catalyzes the electroreduction of carbon dioxide into hydrocarbon fuels. *Energy Environ. Sci.* **3**, 1311–1315 (2010).
- Peterson, A. A. & Nørskov, J. K. Activity descriptors for CO₂ electroreduction to methane on transition-metal catalysts. *J. Phys. Chem. Lett.* **3**, 251–258 (2012).
- Zhang, X. et al. Highly selective and active CO₂ reduction electrocatalysts based on cobalt phthalocyanine/carbon nanotube hybrid structures. *Nat. Commun.* **8**, 14675 (2017).
- Varela, A. S. et al. Metal-doped nitrogenated carbon as an efficient catalyst for direct CO₂ electroreduction to CO and hydrocarbons. *Angew. Chem. Int. Ed.* **54**, 10758–10762 (2015).
- Shen, W. et al. Electrocatalytic reduction of carbon dioxide to carbon monoxide and methane at an immobilized cobalt protoporphyrin. *Nat. Commun.* **6**, 8177 (2015).
- Wu, Y. et al. Electroreduction of CO₂ catalyzed by a heterogenized Zn-porphyrin complex with a redox-innocent metal center. *ACS Cent. Sci.* **3**, 847–852 (2017).
- Pan, Y. et al. Design of single-atom Co-N₅ catalytic site: a robust electrocatalyst for CO₂ reduction with nearly 100% CO selectivity and remarkable stability. *J. Am. Chem. Soc.* **140**, 4218–4221 (2018).
- Trippkovic, V. et al. Electrochemical CO₂ and CO reduction on metal-functionalized porphyrin-like graphene. *J. Phys. Chem. C* **117**, 9187–9195 (2013).
- Ju, W. et al. Understanding activity and selectivity of metal-nitrogen-doped carbon catalysts for electrochemical reduction of CO₂. *Nat. Commun.* **8**, 944 (2017).
- Dhanasekaran, T., Grodkowski, J., Neta, P., Hambright, P. & Fujita, E. p-Terphenyl-sensitized photoreduction of CO₂ with cobalt and iron porphyrins. Interaction between CO and reduced metalloporphyrins. *J. Phys. Chem. A* **103**, 7742–7748 (1999).
- Fujita, E., Creutz, C., Sutin, N. & Szalda, D. J. Carbon dioxide activation by cobalt(II) macrocycles: factors affecting carbon dioxide and carbon monoxide binding. *J. Am. Chem. Soc.* **113**, 343–353 (1991).
- Rao, H., Schmidt, L. C., Bonin, J. & Robert, M. Visible-light-driven methane formation from CO₂ with a molecular iron catalyst. *Nature* **548**, 74 (2017).
- Leonard, N. et al. The chemical identity, state and structure of catalytically active centers during the electrochemical CO₂ reduction on porous Fe-nitrogen-carbon (Fe-N-C) materials. *Chem. Sci.* **9**, 5064–5073 (2018).
- Zhang, H., Li, J., Cheng, M.-J. & Lu, Q. CO Electroreduction: current development and understanding of Cu-based catalysts. *ACS Catal.* **9**, 49–65 (2018).
- Solis, B. H., Maher, A. G., Dogutan, D. K., Nocera, D. G. & Hammes-Schiffer, S. Nickel phlorin intermediate formed by proton-coupled electron transfer in hydrogen evolution mechanism. *Proc. Natl Acad. Sci. USA* **113**, 485–492 (2016).
- Jiang, J. et al. Unusual stability of a bacteriochlorin electrocatalyst under reductive conditions. A case study on CO₂ conversion to CO. *ACS Catal.* **8**, 10131–10136 (2018).
- Grodkowski, J. et al. Reduction of cobalt and iron phthalocyanines and the role of the reduced species in catalyzed photoreduction of CO₂. *J. Phys. Chem. A* **104**, 11332–11339 (2000).
- Lieber, C. M. & Lewis, N. S. Catalytic reduction of carbon dioxide at carbon electrodes modified with cobalt phthalocyanine. *J. Am. Chem. Soc.* **106**, 5033–5034 (1984).
- Zhang, Z. et al. Reaction mechanisms of well-defined metal-N₄ sites in electrocatalytic CO₂ reduction. *Angew. Chem. Int. Ed.* **57**, 16339–16342 (2018).
- Kapusta, S. & Hackerman, N. Carbon dioxide reduction at a metal phthalocyanine catalyzed carbon electrode. *J. Electrochem. Soc.* **131**, 1511–1514 (1984).
- Boutin, E. et al. Aqueous electrochemical reduction of carbon dioxide and carbon monoxide into methanol with cobalt phthalocyanine. *Angew. Chem. Int. Ed.* **58**, 16172 (2019).
- Aoi, S., Mase, K., Ohkubo, K. & Fukuzumi, S. Selective electrochemical reduction of CO₂ to CO with a cobalt chlorin complex adsorbed on multi-walled carbon nanotubes in water. *Chem. Commun.* **51**, 10226–10228 (2015).
- Hu, X.-M., Rønne, M. H., Pedersen, S. U., Skrydstrup, T. & Daasbjerg, K. Enhanced catalytic activity of cobalt porphyrin in CO₂ electroreduction upon immobilization on carbon materials. *Angew. Chem. Int. Ed.* **56**, 6468–6472 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Materials

All chemicals were purchased and used as received without further purification unless otherwise stated. CO₂ (99.99%), CO (99.3%) and N₂ (99.999%) were purchased from Airgas. KHCO₃ (99.7%) was purchased from Sigma Aldrich. Multi-wall CNTs were purchased from C-Nano (product number FT9100). CoPc, NiPc and FePc were purchased from Alfa Aesar. Deionized water used in all experiments was purified through a Milli-Q reference water-purification system to reach a resistivity of 18.2 MΩ cm (at 25 °C).

Characterization

SEM images were taken using a Hitachi SU8230 field-emission SEM microscope. UV-Vis absorption measurements were carried out with a Shimadzu UV-2600 UV-Vis spectrophotometer. STEM imaging was performed using a double Cs-corrected FEI Themis G2 microscope at 60 kV with a Super-X EDS detector. Inductively coupled plasma mass spectrometry (ICP-MS) was performed with an Agilent Technologies 7700 series instrument. ¹H and ¹³C NMR spectra were recorded using an Agilent 400-MHz NMR instrument.

Synthesis of CoPc–NH₂

The synthesis procedure was adapted from previous publications^{29,30}. 4-Nitrophthalonitrile (10 mmol, 1.73 g), CoCl₂·6H₂O (2.5 mmol, 0.60 g), urea (80 mmol, 4.80 g) and a catalytic amount of (NH₄)₆Mo₇O₂₄·4H₂O were first well mixed by grinding in an agate mortar. Then, the mixture was reacted in the solid state at 170 °C for 5 h under Ar atmosphere. The resulting product was stirred at 90 °C for 1 h in HCl (1 M, 200 ml). The solid was filtered and then stirred at 90 °C for 1 h in NaOH (1 M, 200 ml). The crude product was filtered, washed with water, dried in vacuum and then purified by Soxhlet extraction with methanol. The resulting solid was then dissolved in DMF and filtered. The DMF solution was evaporated under vacuum to afford the CoPc–NO₂ compound as a dark-green solid (0.83 g, 44%). The synthesized CoPc–NO₂ (0.75 g, 1 mmol), Na₂S·9H₂O (4.8 g, 20 mmol), 1 ml of deionized water and 25 ml of DMF were mixed in a three-necked round-bottom flask, and the mixture was stirred at 60 °C overnight under Ar atmosphere. After that, the solution was evaporated under vacuum, and the obtained solid was washed with water and then boiled in 100 ml of 5 wt% aqueous NaOH solution. Subsequently, the precipitate was filtered and washed with water. The resulting solid was poured into 250 ml of water while stirring, and 1 M HCl was added to adjust the pH to 5. The mixture was filtered to remove the undissolved side products. The pH of the filtered solution was then adjusted to 8 by adding 1 M KOH, and the resulting solution was boiled. The precipitate was collected by filtration, washed with water and methanol and then dried in vacuum to afford the target CoPc–NH₂ compound as a dark-green solid (0.52 g, 82%). High-resolution mass spectrometry gave a mass-to-charge value of 631.12435 (Extended Data Fig. 9a). UV-Vis spectroscopy²⁹ gave wavelengths of maximum absorbance λ_{max} (in 15 M H₂SO₄) of 210 nm, 299 nm, 380 nm and 739 nm (Extended Data Fig. 9b).

Preparation of MPc/CNT hybrid materials

As-received CNTs were first calcined at 500 °C in air for 5 h. After cooling to room temperature, the CNTs were transferred into a 5 wt% HCl aqueous solution and sonicated for 30 min. The purified CNTs were collected by filtration and washed extensively with deionized water. 30 mg of the purified CNTs was subsequently dispersed in 30 ml of DMF using sonication (XM-300UHP, 600 W/10 L, 40 KHz). Then, an appropriate amount of MPc (1.5 mg CoPc, 1.2 mg CoPc–NH₂, 1.6 mg FePc or 1.1 mg NiPc) dissolved in DMF was added to the CNT suspension. The mixture was sonicated for 30 min to obtain a well mixed suspension, which was further stirred at room temperature for 20 h. Subsequently, the mixture was centrifuged and the precipitate was washed with DMF and ethanol. Finally, the precipitate was lyophilized to yield the final product. The

weight percentage of metal in the hybrid material was ~0.27% for all MPc/CNT materials, as confirmed by ICP-MS measurements.

Electrode preparation

Catalyst ink was prepared by dispersing 2 mg of hybrid materials (CoPc/CNT, CoPc–NH₂/CNT, FePc/CNT or NiPc/CNT) in 2 ml of ethanol with 6 μl of 5 wt% Nafion solution, followed by sonication for 1 h. 200 μl of the ink was then drop-casted onto a 3 × 0.5 cm² polytetrafluoroethylene-treated carbon fibre paper (Toray O30, Fuel Cell Store) to cover an area of 0.5 × 1 cm² (catalyst mass loading, 0.4 mg cm^{−2}). The prepared electrodes were fully dried using an infrared lamp. The physical-mixture electrodes used for electrochemical and UV-Vis studies were prepared in the same way, except that 1 mg of CoPc (or CoPc–NH₂) was mixed with 1 mg of carbon material (purified CNTs, Vulcan XC72 or Ketjenblack) before the addition of ethanol and Nafion and that the total mass loading was 0.4 mg cm^{−2} on a 1 × 1 cm² area. The free CoPc electrode was prepared by drop-casting a 0.05 mg ml^{−1} CoPc/DMF solution onto the carbon fibre paper to cover an area of 1 × 0.5 cm² on a heating plate held at 130 °C to reach a final mass loading of 0.1 mg cm^{−2}. For cyclic-voltammetry measurements, 7.5 μl of CoPc/CNT or CoPc–NH₂/CNT ink was deposited on a well polished glassy carbon electrode (electrode diameter, 4 mm; mass loading, 0.06 mg cm^{−2}).

Electrolyte purification

500 ml of a 0.1 M KHCO₃ aqueous solution was purified by a two-step electrolysis using two 10 × 5 cm² high-purity Ti foil (99.99%) electrodes in a two-electrode setup. The first electrolytic step was conducted at 2.5 V until the current decreased to 150 μA. The second step was performed at a constant current of 150 μA for at least 20 h. During the electrolysis, the solution was magnetically stirred. The Ti electrodes were removed from the solution before the electrolysis was terminated to avoid re-dissolution of electrodeposited impurities into the solution.

Electrochemical measurements

Electrochemical experiments were performed using a Bio-Logic VMP3 Multi Potentiostat and a custom-designed gas-tight two-compartment electrochemical cell. The graphite rod counter-electrodes were purchased from Sigma Aldrich and the Ag/AgCl reference electrodes (0.199 V versus SHE) were purchased from Pine Research Instrumentation. The cathode and anode compartments were separated by an anion-exchange membrane (Selemon DSV). Each compartment contained 12 ml of electrolyte and ~18 ml of gas headspace. For all experiments, the pre-purified 0.1 M KHCO₃ was used as the electrolyte. Before each measurement, the electrolyte was pre-saturated with CO₂, N₂ or CO by bubbling the gas for at least 15 min. Gas was continuously bubbled into the electrolyte during electrolysis (or flowed into the headspace during cyclic-voltammetry measurements) at a flow rate of 20 standard cubic centimetres per minute. Before the start of each electrolysis, the Ohmic drop between the working electrode and the reference electrode was determined using potentiostatic electrochemical impedance spectroscopy at −0.5 V versus Ag/AgCl between 200 kHz and 1 Hz with an amplitude of 10 mV. The resistance was then determined by the intersection of the curve with the real axis of the Nyquist plot. Correction for internal resistance was performed after electrolysis for all measurements except for the electrodes used for the UV-Vis studies, where the internal resistance drop was compensated during the electrochemical measurement. Current densities were calculated on the basis of the catalyst-covered geometric area of the working electrode. All potentials (V) were converted to the RHE scale using the following formula: $V_{\text{RHE}} = V_{\text{Ag/AgCl}} + (0.199 \text{ V}) + (0.0592 \text{ V}) \times \text{pH}$.

Product quantification

The gas products of electrocatalysis were analysed using a gas chromatography system (SRI Multiple Gas Analyzer #5) equipped with a flame ionization detector and a thermal conductivity detector. High-purity

N₂ was used as the carrier gas. The peak areas of the products (H₂ and CO) were converted to gas volumes using calibration curves that were obtained using a standard gas diluted with CO₂ to different concentrations. The liquid products were quantified after electrocatalysis using ¹H NMR spectroscopy with solvent (H₂O) suppression. 400 µl of electrolyte was mixed with 100 µl of a solution of 10 mM dimethyl sulfoxide (DMSO) and 50 mM phenol in D₂O as internal standards for the ¹H NMR analysis. The concentration of MeOH was calculated using the ratio of the area of the MeOH peak (at a chemical shift of 3.31 ppm) to that of the DMSO internal standard (see Extended Data Fig. 2e, f for details). ¹³C NMR spectroscopy was performed for the samples to further verify the presence of MeOH. The FEs for the gas-phase products were average values from three measurements in a single electrolysis experiment. The FE_{MeOH} values at the optimal potentials (shown with error bars in Figs. 2 and 4) were averages from three different 1-h electrolyses, whereas the MeOH FEs at all other potentials were measured once.

UV-Vis characterization of CoPc, CoPc–NH₂ and chemically reduced CoPc

UV-Vis absorption spectra of pristine CoPc and CoPc–NH₂ were taken using their 0.1 mg ml^{−1} DMF solutions. A used CoPc/CNT or CoPc–NH₂/CNT electrode was taken out of the electrolyte solution after the electrolysis and quickly dipped into a vial containing 3 ml of deoxygenated DMF. The vial was immediately sealed, gently sonicated for 10 s and then kept still for another 60 min, after which the supernatant was used for UV-Vis measurement. The chemical reduction of CoPc was done by

mixing equal volumes of 0.1 mg ml^{−1} CoPc dissolved in deoxygenated DMF and 1 mg ml^{−1} NaBH₄ in deoxygenated DMF.

Data availability

Data supporting the findings of this study are available from the corresponding authors upon reasonable request. Source data for Figs. 2, 4 are provided with the paper.

29. Achar, B. N. & Lokesh, K. S. Studies on tetra-amine phthalocyanines. *J. Organomet. Chem.* **689**, 3357–3361 (2004).
30. Ding, X. & Han, B.-H. Metallophthalocyanine-based conjugated microporous polymers as highly efficient photosensitizers for singlet oxygen generation. *Angew. Chem. Int. Ed.* **54**, 6536–6539 (2015).

Acknowledgements This work was supported by the US National Science Foundation (grant CHE-1651717). X.L. acknowledges a Croucher Fellowship for Postdoctoral Research. Z.J. and Y.L. acknowledge financial support by Shenzhen fundamental research funding (JCYJ20160608140827794) and the Guangdong Provincial Key Laboratory (2018B030322001). The authors thank Q. Wang and M. Gu (Southern University of Science and Technology) for assistance with STEM imaging.

Author contributions Y.W. and H.W. conceived the project and designed the experiments. Z.J. and Y.L. synthesized the catalyst materials and performed the structural characterizations. Y.W. and X.L. performed the electrocatalytic studies. Y.W., Y.L. and H.W. analysed the data. Y.W. and H.W. wrote the manuscript. Y.L. and H.W. supervised the project.

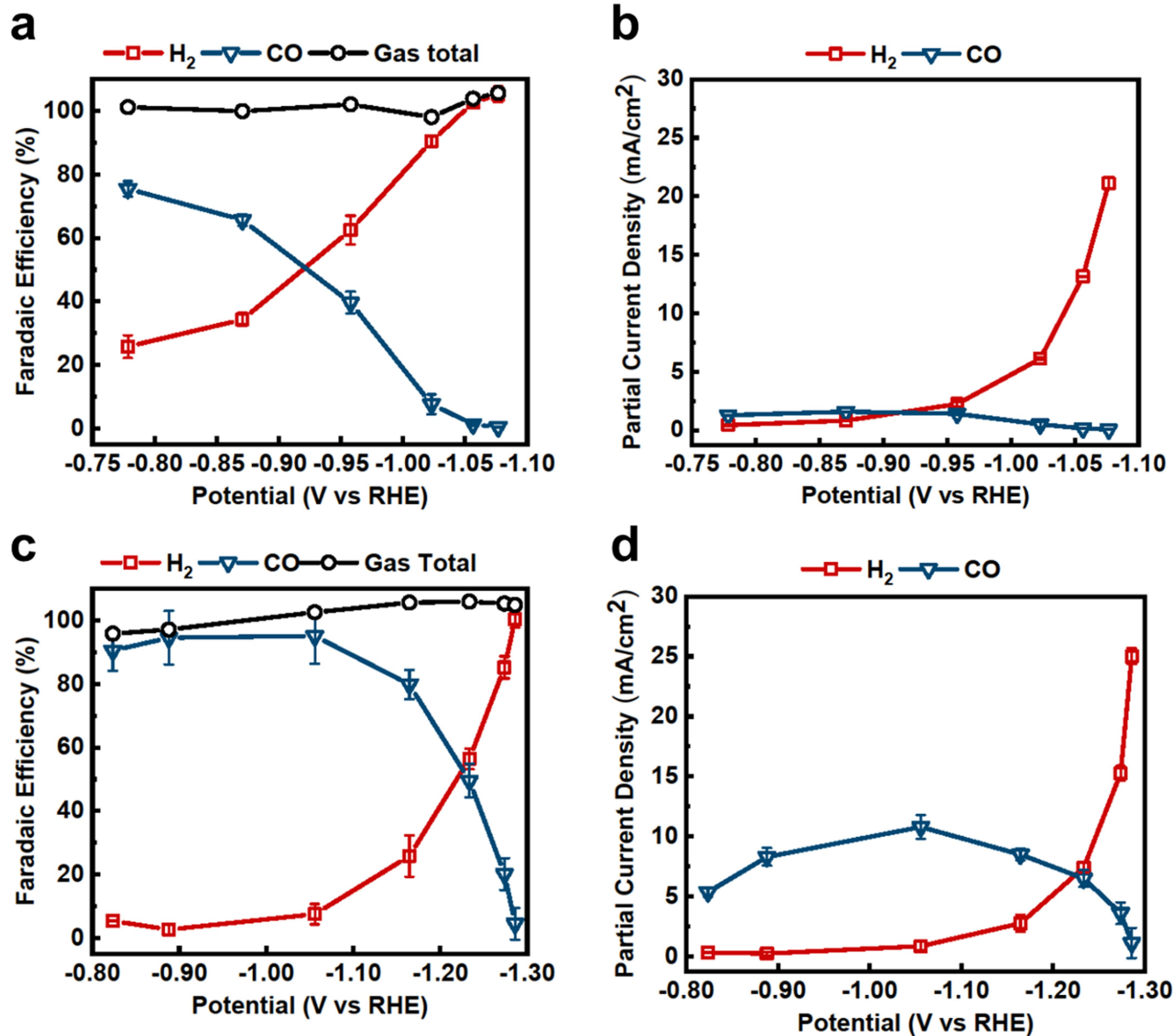
Competing interests The authors declare no competing interests.

Additional information

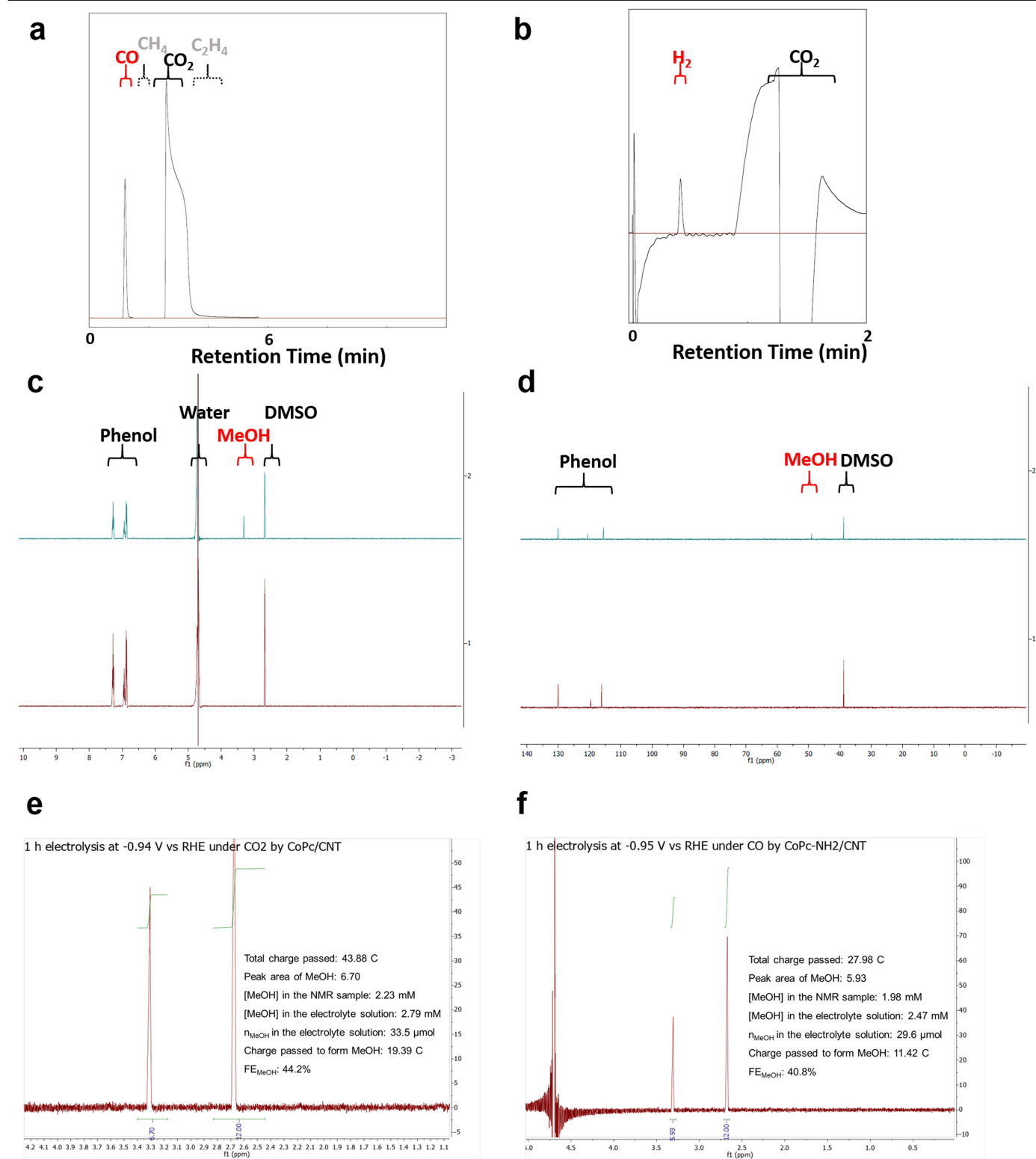
Correspondence and requests for materials should be addressed to H.W. or Y.L.

Peer review information Nature thanks Robert Francke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



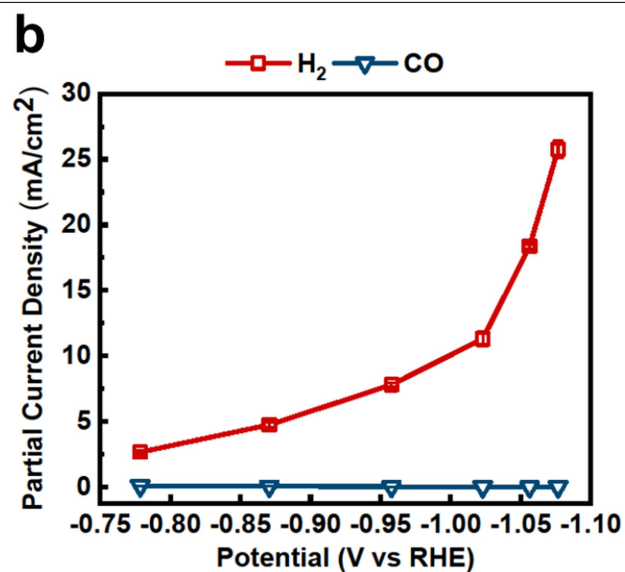
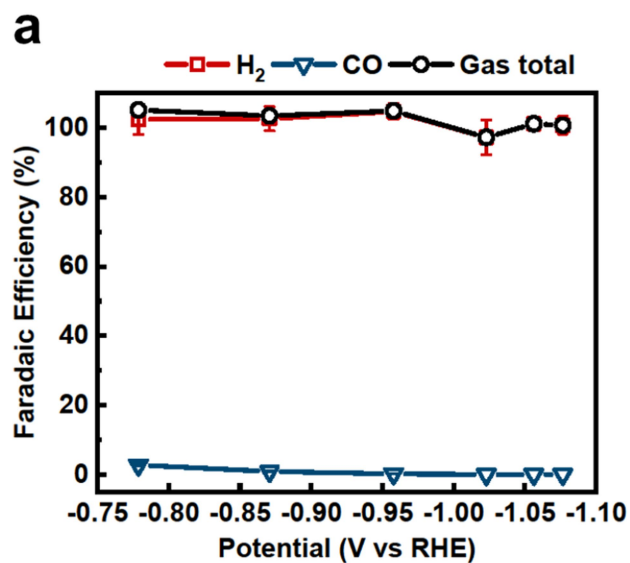
Extended Data Fig. 1 | Catalytic properties of FePc/CNT and NiPc/CNT. **a–d**, Potential-dependent catalytic performance of CO₂ electroreduction by FePc/CNT (**a, b**) and NiPc/CNT (**c, d**).



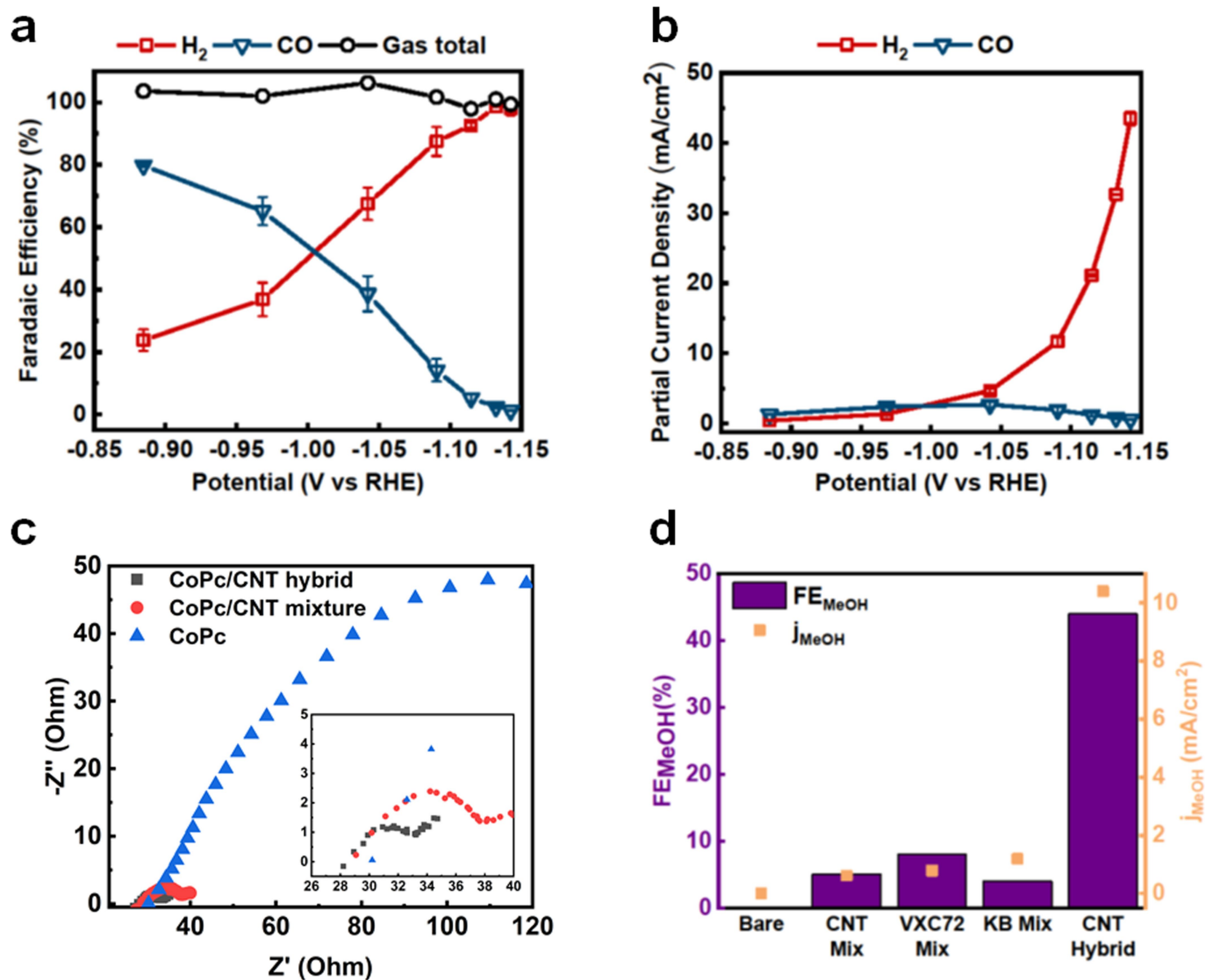
Extended Data Fig. 2 | Product identification and quantification.

a, b, Typical gas chromatography diagrams from the flame ionization detector (a) and the thermal conductivity detector (b), showing the presence of CO and H₂ (marked in red) and the absence of other common gas products (marked in grey). **c, d**, Typical ¹H NMR (c) and ¹³C NMR (d) spectra of a liquid sample after CO₂ electroreduction electrolysis (green traces) versus a blank 0.1 M KHCO₃

solution (red traces). **e, f**, Representative ¹H NMR spectra of liquid samples after 1 h of CO₂ reduction electrolysis catalysed by CoPc/CNT at -0.94 V versus RHE (e) and 1 h of CO reduction electrolysis catalysed by CoPc-NH₂/CNT at -0.95 V versus RHE (f). The detailed information used to determine the FE of MeOH production is given in the diagrams.

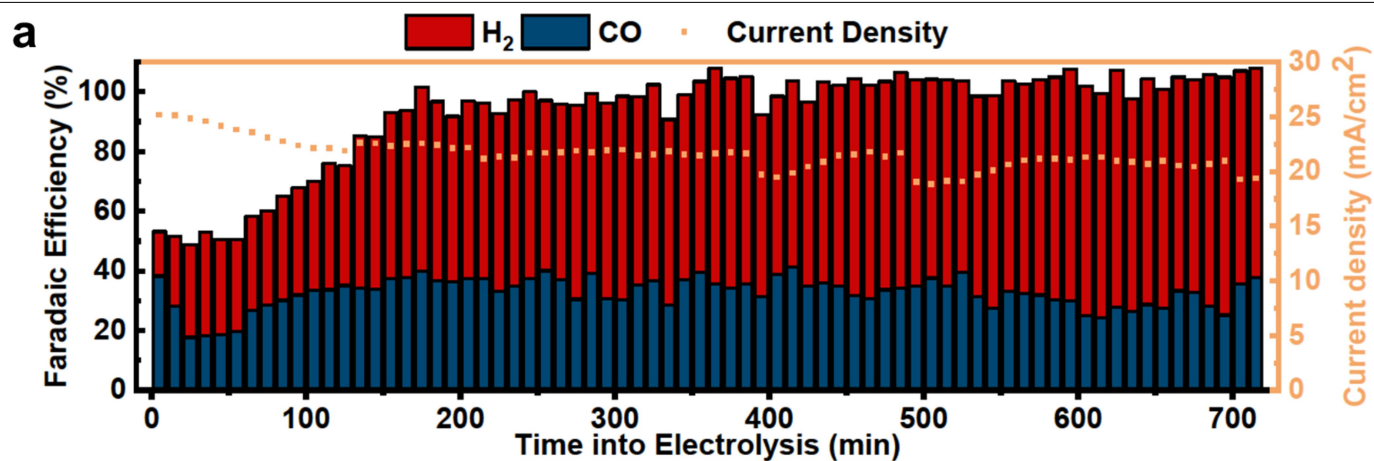


Extended Data Fig. 3 | Catalytic performance of CoPc/CNT under N_2 . a, b, Potential-dependent product selectivity (a) and partial current density (b) for CO_2 electroreduction catalysed by CoPc/CNT under N_2 .

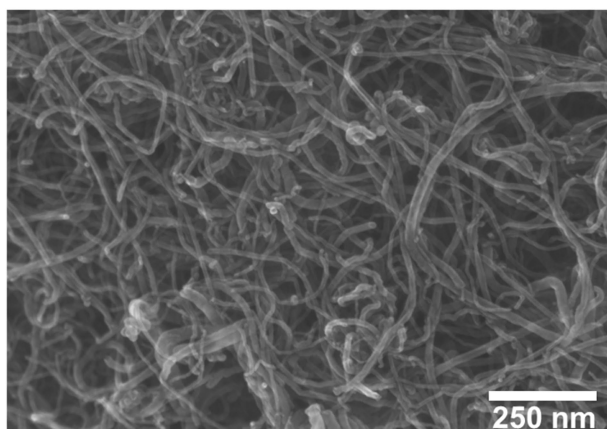


Extended Data Fig. 4 | Effects of carbon support on CO_2 -to-MeOH conversion. **a, b**, Potential-dependent product selectivity (**a**) and partial current density (**b**) for the electroreduction of CO_2 by bare CoPc aggregates on carbon paper. **c**, Nyquist plots of bare CoPc, a physical mixture of CoPc and CNTs and the CoPc/CNT hybrid, measured at -0.94 V versus RHE in a 0.1 M aqueous $KHCO_3$ solution saturated with CO_2 with a scanning frequency range from 200 kHz to 100 Hz and an a.c. amplitude of 5 mV. Z' and Z'' refer to the real

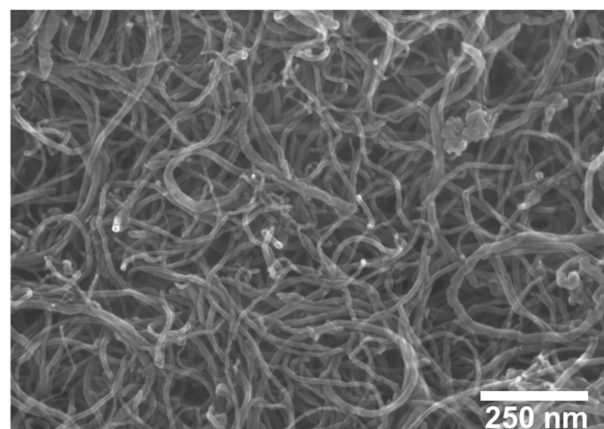
and imaginary parts of impedance, respectively. The inset shows the enlarged area where the traces for the CoPc/CNT hybrid and the mixture of CoPc and CNTs are clearer. **d**, Selectivity (FE_{MeOH}) and partial current density (j_{MeOH}) for CO_2 reduction to MeOH catalysed by bare CoPc ('Bare'), the CoPc/CNT physical mixture ('CNT Mix'), the CoPc/Vulcan XC72 mixture ('VXC72 Mix'), the CoPc/Ketjenblack mixture ('KB Mix') and the CoPc/CNT hybrid ('CNT Hybrid') at -0.94 V versus RHE.



b

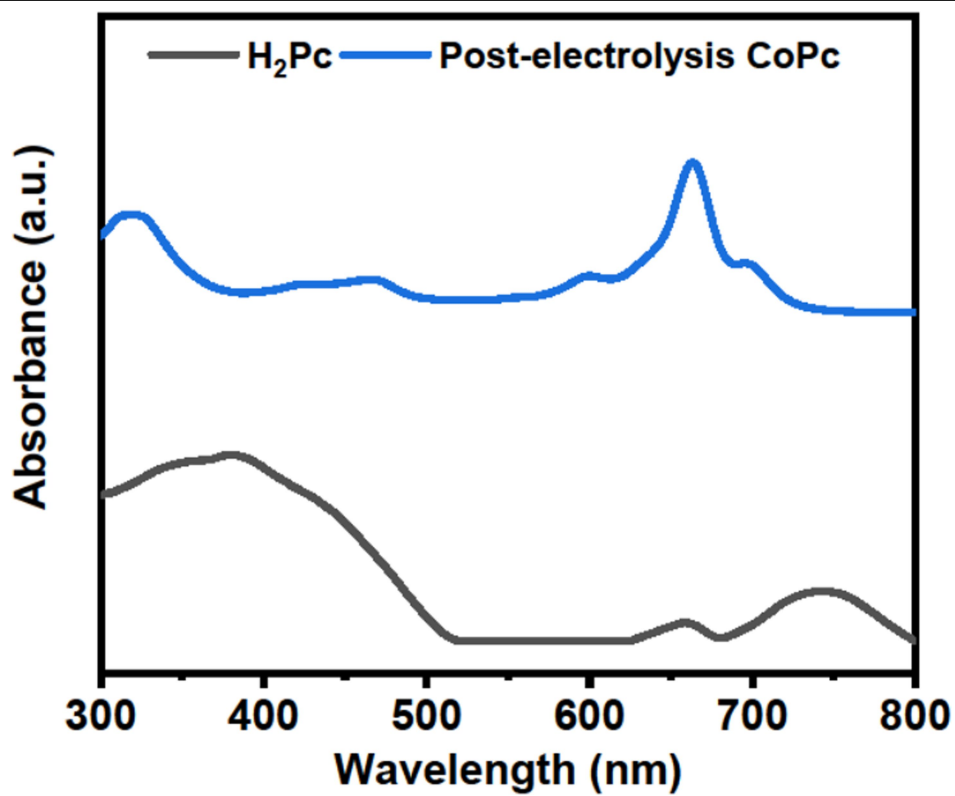


c

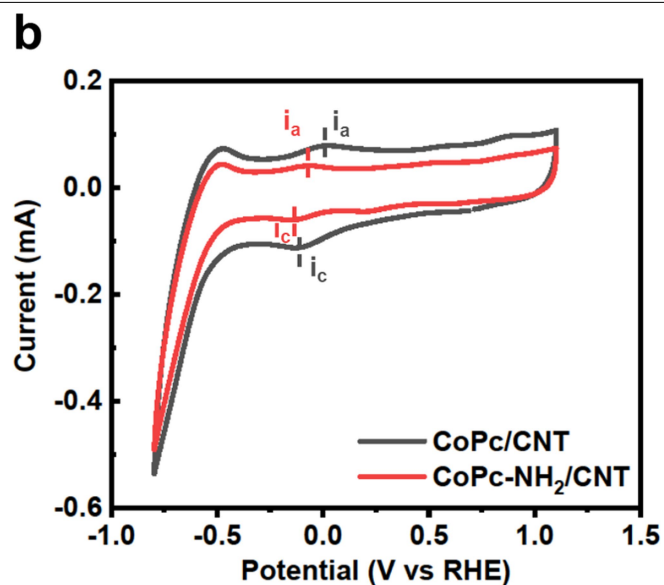
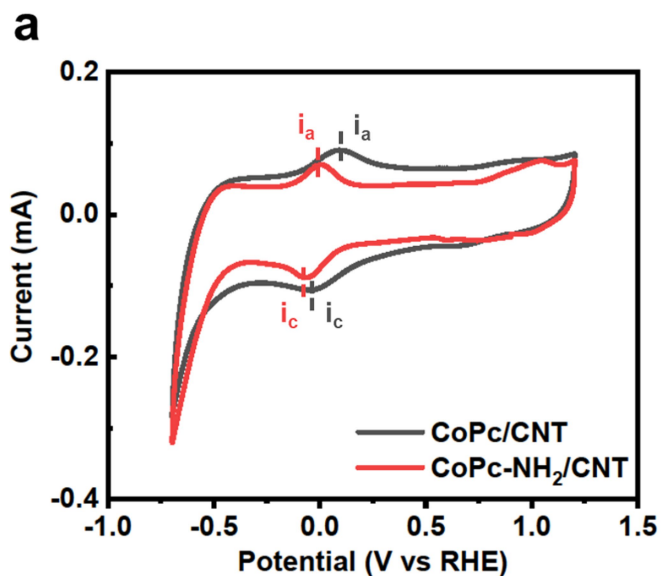


Extended Data Fig. 5 | Long-term electrocatalytic performance of CoPc/CNT. a, Gas product selectivity and total current density in a 12-h electrolysis of CO₂ reduction catalysed by CoPc/CNT at -0.94 V versus RHE. **b, c,** SEM images

of an as-deposited CoPc/CNT electrode before (**b**) and after (**c**) catalysing CO₂ reduction electrolysis at -0.94 V versus RHE for 12 h.

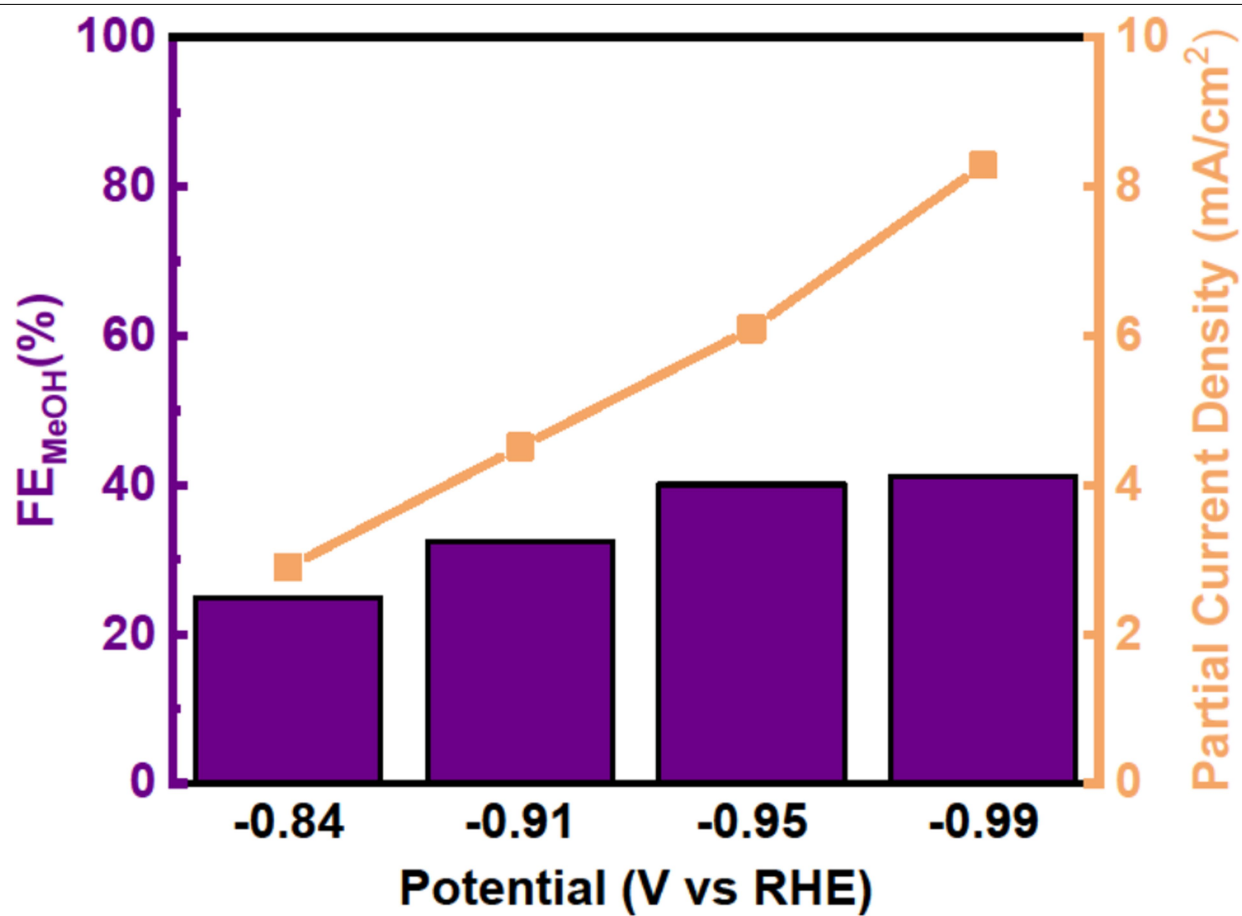


Extended Data Fig. 6 | UV-Vis spectra of post-electrolysis CoPc and free-base phthalocyanine (H₂Pc) in DMF.

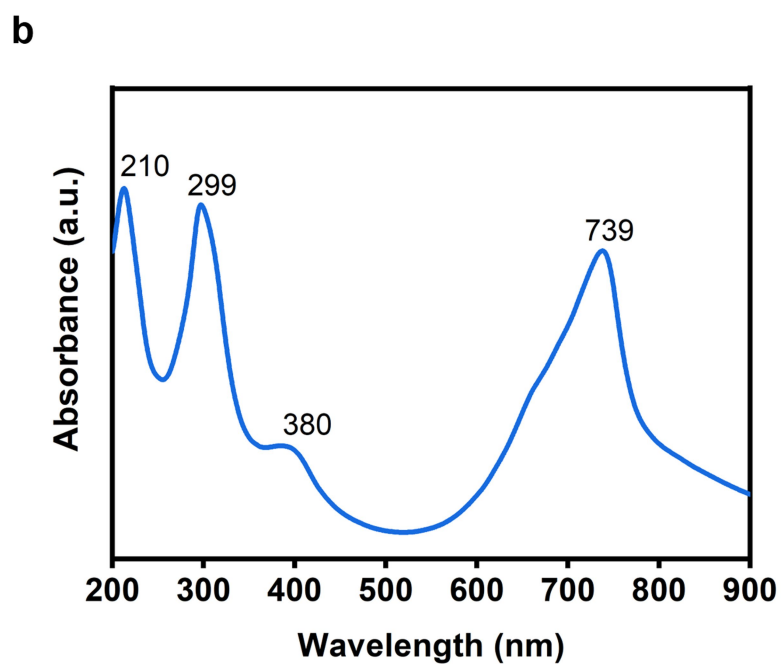
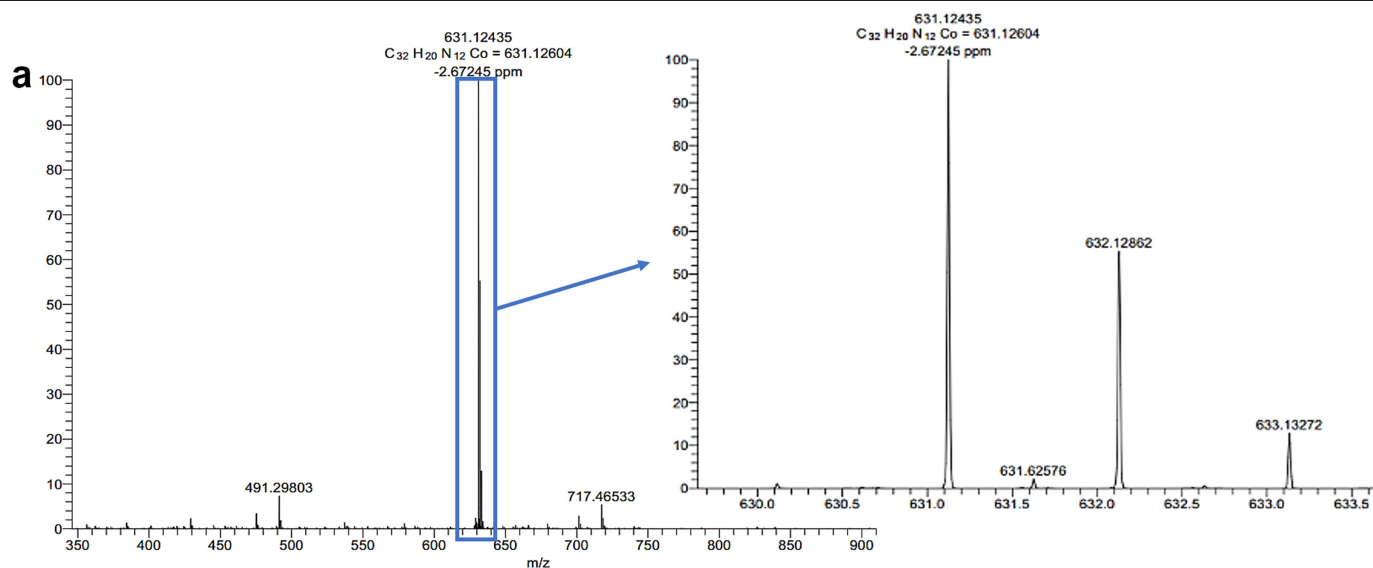


Extended Data Fig. 7 | Cyclic voltammograms of CoPc/CNT and CoPc-NH₂/CNT. a, b, The curves were recorded at a scan rate of 400 mV s⁻¹ in a 0.1 M aqueous KHCO₃ solution under N₂ (a) and CO₂ (b). The most prominent

cathodic (i_c) and anodic (i_a) features of the molecules are labelled. The redox potential of CoPc-NH₂/CNT is more negative than that of CoPc/CNT.

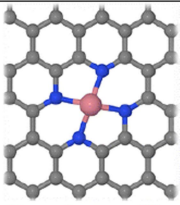
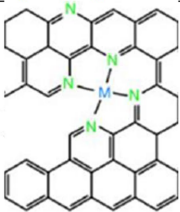
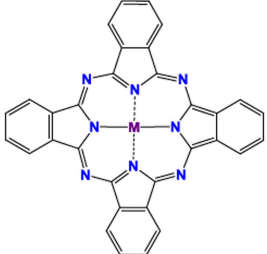


Extended Data Fig. 8 | Catalytic performance of CoPc-NH₂/CNT for electroreduction of CO to MeOH.



Extended Data Fig. 9 | Spectroscopic characterization of CoPc-NH₂. **a**, High-resolution mass spectrometry results for CoPc-NH₂. Calculated for CoC₃₂H₂₀N₁₂: 631.12604; found: 631.12435. **b**, UV-Vis spectrum of CoPc-NH₂ in 15 M H₂SO₄. λ_{max} = 210 nm, 299 nm, 380 nm and 739 nm.

Extended Data Table 1 | Electronic binding energy of CO to the metal centre of M-N₄ motifs (where M = Ni, Co or Fe)

Metal center	Oxidation state	Peripheral structure	$E_B(\text{CO})$ (eV)	Span of $E_B(\text{CO})$ from Fe to Ni	Reference
Fe	+2	 Graphene-like porphyrin	-1.13	1.19	J. Phys. Chem. C 2013, 117, 9187
Co	+2		-0.52		
Ni	+2		0.06		
Fe	+2	 Nitrogen-doped carbon scaffold	-1.27	1.23	Nat. Commun. 2017, 8, 944
Co	+2		-0.60		
Ni	+2		-0.04		
Fe	+2	 Phthalocyanine ligand	-1.24	1.19	Angew. Chem. Int. Ed. 2018, 57, 16339
Co	+2		-0.62		
Ni	+2		-0.05		

The $E_B(\text{CO})$ values were obtained by converting the corresponding DFT-computed free energies of CO binding ($G_B(\text{CO})$) reported in the literature^{13,14,24}. The conversion was done by subtracting a 0.5-eV energy term^{6,7}.

Extended Data Table 2 | Performance comparison between our catalyst system and previously reported transition-metal complex electrocatalysts for CO₂-to-MeOH conversion

Catalytic system	j _{MeOH} (mA/cm ²)	Optimal FE _{MeOH} (%)	Electrolyte solution	Potential	TOF	TON ^c	Comments	Reference
CoPc/CNT	10.6 (1 h)	44 (1 h)	0.1 M KHCO ₃	-0.94 V vs RHE	1.05	~3800		This work
CoPc-NH ₂ /CNT	10.2 (1 h)	32 (1 h)	0.1 M KHCO ₃	-1.00 V vs RHE	1.01	~3600		This work
CoPc-NH ₂ /CNT	8.9 (12 h)	28 (12 h)	0.1 M KHCO ₃	-1.00 V vs RHE	0.88	~38000		This work
Mixture of CoPc and CNT	0.03 (3 h)	0.3 (3 h)	0.5 M KHCO ₃	-0.88 V	~0.03 ^b	44		10.1002/anie.201909257
CoPc	~0.05	<5	pH = 3 acid solution	-1.2 V ~ -1.4 V vs SCE ^d	~0.005 ^b	N/A		J. Electrochem. Soc. 1984, 131, 1511-1514
[(phen) ₂ Ru(dppz)] ²⁺	~0.06 (6 h) ^a	~100 (6 h)	DMF with 0.30 mM [(phen) ₂ Ru(dppz)] ²⁺ , 0.1 M TBAPF ₆ and 1 M H ₂ O	-0.6 V vs Ag/AgCl	~4 × 10 ⁻⁴ ^b	0.92	Based on pyridine	Inorg. Chem. 2014, 53, 6544
Pt/ Polyaniline/ Prussian blue/ 2-hydroxy-1-nitronaphthalene-3,6-disulphonatocobalt(II)	~0.0006 (24 h) ^a	~0.2 (24 h)	0.5 M KCl with HCl (pH = 2.0)	-0.6 V vs SCE ^d	~3 × 10 ⁻⁶ ^b	~0.3		J. Chem. Soc., Chem. Commun., 1993, 0, 20
Pt/ metal tetraphenylporphyrin/ aquapentacyanoferrate(II) / methanol	~0.03 (5 h) ^a	15.1 (5 h)	0.1 M KCl with 15 mM 2-hydroxy-1-nitronaphthalene-3,6-disulphonatocobalt(II), 20 mM MeOH	-0.5 V vs SCE ^d	~0.001 ^b	~20	A much larger amount of MeOH added prior to electrolysis than that produced in the reaction	Journal of Molecular Catalysis, 47 (1988) 51 - 51
Pt (stainless steel)/ Quinone derivatives/ aquapentacyanoferrate(II)/ Methanol	~0.01 (2 h) ^a	70.2 (2 h)	0.1 M KCl with 10 mM aquapentacyanoferrate(II) and 15 mM MeOH	-0.34 V ~ -0.39 V vs SCE ^d	Unable to estimate	N/A		Journal of Molecular Catalysis, 41 (1987) 303 - 311
Pt/ Everitt's salt/ aquapentacyanoferrate(II)/ methanol	~0.02 (10 h) ^a	102 (10 h)	0.1 M KCl with 15 mM aquapentacyanoferrate(II) and 20 mM MeOH	-0.6 V vs SCE ^d	~1 × 10 ⁻⁴ ^b	~4		J. Electroanal. Chem., 220 (1987) 333-337
Pt/ Everitt's salt/ 1-nitroso-2-naphthol-3,6-disulfonic acid metal complex/ methanol	~0.02 (5 h) ^a	83.4 (5 h)	0.1 M KCl with 15 mM 2-naphthol-3,6-disulfonic acid Fe(II) and 20 mM MeOH	-0.6 V vs SCE ^d	~1 × 10 ⁻⁴ ^b	~2		J. Electroanal. Chem., 206 (1986) 209-216
Pt/ 1,2-dihydroxybenzene-3,5-disulphonatoferrate(III) / Ethanol	~0.04 (6 h) ^a	Not available	0.1 M KCl with 20 mM ethanol, 1 mM Fe ³⁺ and 2 mM citron	-1.0 V vs SCE ^d	Unable to estimate	N/A		Journal of Molecular Catalysis, 34 (1986) 67 - 72

^aCalculated using the reported selectivity, total charge passed (or current), duration and geometric area of the working electrode.

^bCalculated using the partial current density and catalyst loading (or amount of catalyst in the solution).

^cTurnover number (TON) in the reported catalysis, calculated on the basis of the turnover frequency (TOF) and measurement duration.

^dSaturated calomel electrode.

Concise asymmetric synthesis of (–)-bilobalide

<https://doi.org/10.1038/s41586-019-1690-5>

Received: 10 June 2019

Accepted: 8 August 2019

Published online: 16 October 2019

Meghan A. Baker^{1,3}, Robert M. Demoret^{1,3}, Masaki Ohtawa^{1,2*} & Ryan A. Shenvi^{1*}

The *Ginkgo biloba* metabolite bilobalide is widely ingested by humans but its effect on the mammalian central nervous system is not fully understood^{1–4}. Antagonism of γ -aminobutyric acid A receptors (GABA_ARs) by bilobalide has been linked to the rescue of cognitive deficits in mouse models of Down syndrome⁵. A lack of convulsant activity coupled with neuroprotective effects have led some to postulate an alternative, unidentified target⁴; however, steric congestion and the instability of bilobalide^{1,2,6} have prevented pull-down of biological targets other than the GABA_ARs. A concise and flexible synthesis of bilobalide would facilitate the development of probes for the identification of potential new targets, analogues with differential selectivity between insect and human GABA_ARs, and stabilized analogues with an enhanced serum half-life⁷. Here we exploit the unusual reactivity of bilobalide to enable a late-stage deep oxidation that symmetrizes the molecular core and enables oxidation states to be embedded in the starting materials. The same overall strategy may be applicable to *G. biloba* congeners, including the ginkgolides—some of which are glycine-receptor-selective antagonists⁸. A chemical synthesis of bilobalide should facilitate the investigation of its biological effects and its therapeutic potential.

The leaves of *Ginkgo biloba* have been used historically as insecticides and helminthicides^{9,10}, and this activity has been attributed to their constituent terpene trilactones, including bilobalide^{11,12} (1, Fig. 1). *Ginkgo* extracts have also been used in traditional Chinese medicine to treat senility, a practice that has penetrated the Western world—although not without controversy, owing to opposing claims of efficacy¹ and serious adverse effects associated with Ginkgo toxin (4-*O*-methylpyridoxine)² or the inhibition of platelet aggregating factor¹³ by ginkgolides. Animal models demonstrate some credible effects on impaired cognition: in a mouse model of Down syndrome (Ts65DN), in which mice show deficits in declarative learning and memory, normalized novel object recognition was exhibited after treatment with pure bilobalide⁵. Rescue of learning and memory is proposed to arise through neuronal excitation by antagonism of GABA_ARs. Unlike the plant metabolite picrotoxinin (2, Fig. 1), bilobalide is not acutely toxic, and unlike the ginkgolides, bilobalide does not affect platelet aggregating factor. Despite their disparate toxicity, bilobalide and picrotoxinin exhibit similar inhibitory potencies at recombinant GABA_ARs, with half-maximum inhibitory concentrations (IC₅₀) of 4.6 μ M and 2.0 μ M, respectively ($\alpha_1\beta_2\gamma_{2L}$ receptor, *Xenopus laevis* oocytes), yet the two compounds show differential inhibition of GABA_AR-positive modulators¹⁴.

In addition to incomplete approaches^{15–17}, two previous syntheses of bilobalide have been completed (24 steps enantioselective^{18,19}; 17 steps racemic²⁰), both of which established the cyclopentane core with efficiency but required 8–11 subsequent redox steps to reach the target compound. Guided by these challenges, we realized that a single oxidation transform might reduce synthetic complexity by unmasking a pseudosymmetric fused dilactone, ultimately leading

to a symmetric starting material (Fig. 1). However, late-stage installation of the deep C10 hydroxyl (Fig. 1, highlighted in red) presented a problem. Neither hydrogen of its precursor inner lactone (shown in green) seemed accessible, whereas a hydrogen of the outer lactone (shown in blue) resided at the surface of the bowl-like scaffold. Here we utilize unexpected properties of the bilobalide architecture in a concise synthesis of (–)-*des*-hydroxybilobalide (>99% enantiomeric excess (e.e.), 5, Fig. 1), which relies on stereocontrol transmitted from an unusual oxetane acetal. A late-stage oxidation is rendered regioselective using skeletal rearrangement and acidification, completing the synthesis of (–)-bilobalide in a single additional step.

The synthesis commenced with a methodological challenge: an asymmetric Reformatsky reaction between **6a** and **6b** (which are produced in two steps and one step respectively, see Supplementary Information section 3). Reformatsky conditions proved necessary owing to the tendency of **7** to undergo retro-aldol cleavage under basic conditions, whereas zinc, chromium and samarium alkoxides were stable at –78 °C. To our knowledge, there have been no examples of catalytic enantio- and diastereoselective zinc Reformatsky reactions²¹, nor the use of simple, chiral L-type bisoxazoline (BOX) ligands. A previous study demonstrated the use of related, electron-rich hemiaminals for the control of single stereocentres²², and the simplicity of these conditions provided a foundation to explore. After a ligand screen, we found that a combination of diethylzinc and indabox (10 mol% **A**) provided secondary alcohol **7** in 97:3 enantiomeric ratio (e.r.) in favour of *syn*-diastereomer **7** (2.3:1). The combined yield of both diastereomers was determined to be 64% by NMR, and the crude reaction mixture could be carried forward efficiently (purification by chromatography

¹Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA. ²Present address: Graduate School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan. ³These authors contributed equally: Meghan A. Baker, Robert M. Demoret. *e-mail: ohtawam@pharm.kitasato-u.ac.jp; rshenvi@scripps.edu

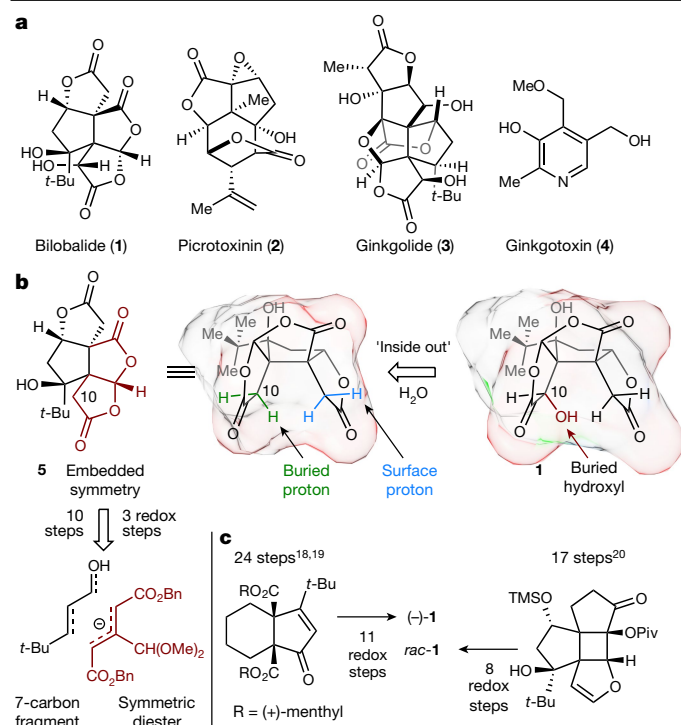


Fig. 1 | Congeners and design considerations. **a**, Plant metabolites that show activity in the human nervous system. **b**, Late-stage excision of a buried C–H bond in bilobalide enables oxygen atoms to be placed in a symmetric starting material. **c**, Previous work iteratively changed oxidation states.

on silica led to a loss of material via a retro-aldol reaction). The doubly activated bromide **6a** may uniquely enable the use of simple BOX ligands²¹.

A Giese-type 5-*exo*-trig cyclization of **7** occurred with high regio- and diastereoselectivity (20:1 diastereoselective ratio (d.r.); the 6-*endo*-trig product was not detected) to form the quaternary carbon of cyclopentene **8**. The material was purified by recrystallization to >99% e.e. and a 21% yield was achieved over two steps; the relative and absolute stereochemistry of the product were confirmed by X-ray crystallography (see Supplementary Information section 5).

Previous syntheses of **1** installed the extremely hindered *tert*-alkyl, bis-*neo*-pentyl C8 hydroxyl either by late-stage dihydroxylation with stoichiometric osmium tetroxide (23 °C, 12 h)¹⁸; or by early-stage nitrile anion addition to a *tert*-butyl ketone, which rendered the synthesis racemic²⁰. More recently, the Drago–Mukaiyama hydration has emerged as an effective method by which to hydrate hindered alkenes in a chemoselective manner through an outer-sphere, metal-hydride hydrogen-atom-transfer mechanism²³. In this case, the use of standard conditions resulted in a low yield of **9** with no diastereoselectivity (Fig. 2b, entry 4). We recently discovered that the kinetically relevant reductant in many Mukaiyama reactions is an alkoxy silane (for example, Ph(*i*-PrO)SiH₂) that is formed in situ by silane alcoholysis²⁴. This custom, commercially available silane enabled us to screen a diverse range of solvents, and we identified a correlation between solvent polarity and diastereoselectivity, possibly due to internal hydrogen bonding (Fig. 2b). Variation of the ligands on the metal catalyst had no effect on the diastereomeric ratio of the product. The use of *tert*-butyl methyl ether favoured the wrong (*S*)-C8 diastereomer of the alcohol (which cyclized to a lactone), whereas methylcyclohexane reversed this stereoselectivity to favour the (*R*)-C8 diastereomer **9** in a 3:1 ratio.

Formation of the fourth contiguous, fully substituted carbon atom by alkylation was frustrated by the dehydration of **9** under basic conditions or by a preference for the wrong C5 stereoisomer.

However, treatment of **9** with strong acid, such as *p*-toluenesulfonic acid (PTSA) (Fig. 2c) led to unexpectedly stable oxetane acetals **10** (*endo*-OMe) and **11** (*exo*-OMe). Stabilization of **10** and **11** is possibly driven by steric compression of the alcohol and acetal carbons—a ‘corset effect’ that increases the energy barrier to ring-opening of strained molecules such as tetrahydranes²⁵. Only the minor *endo*-isomer **10** could be carried forward: the major *exo*-OMe diastereomer **11** dehydrated under basic conditions and its epimerization to **10** was unsuccessful. However, early racemic route scouting had provided a possible way forward. Screening a library of scalemic binolphosphoric acids had been expected to improve diastereoselectivity, but this effort yielded unsatisfactory results: a 1.7:1 preference for **10** using catalyst **B**. Analysis by chiral chromatography indicated that a moderate parallel kinetic resolution of (*rac*)-**9** had occurred: each diastereomer possessed opposite enantiomeric excess (**10**, 39:61 e.e. compared with **11**, 69:31 e.e.). Accordingly, a single enantiomer of **9**, if matched with the correct enantiomer of chiral acid, would favour formation of the desired *endo*-**10**. Indeed, whereas (–)-**9** reacted with (–)-**B** to favour (1.4:1) *exo*-acetal **11**, (+)-**9** reacted to favour (4.5:1) *endo*-acetal **10**, isolated in 71% yield as a single enantiomer. *endo*-Acetal **10** proved crucial to control the formation of the final quaternary carbon.

Owing to extreme steric hindrance in substrate **10**, the final C–C bond could be established only using an alkyne electrophile. A three-step sequence was run in quick succession owing to intermediate instability; only alkyne **12** could be purified. First, oxidation using 2-iodoxybenzoic acid (IBX) delivered an unstable β-keto ester that could undergo α-alkylation. Second, this mixture of keto–enol tautomers was treated with Waser’s reagent (1-[(trimethylsilyl)ethynyl]-1,2-benziodoxol-3(1*H*)-one; TMS-EBX) and tetra-*n*-butylammonium fluoride (TBAF)²⁶. The *endo*-methoxy oxetane effectively shielded one trajectory of electrophile approach and provided the product as a single diastereomer. By contrast, the *exo*-methoxy oxetane—if carried forward to this step—eliminated the *tert*-alkyl ether and did not undergo alkynylation. Alkynylation before oxetane formation delivered exclusively the incorrect diastereomer. The unstable alkynylation product was reduced with high stereoselectivity using samarium iodide (SmI₂) in a mixture of tetrahydrofuran/water to provide the stable alcohol **12** in 60% yield over three steps²⁷. Traditional hydride reductants produced the opposite diastereomer. Anti-Markovnikov hydration of the alkyne to directly incorporate the northern lactone motif could not be accomplished under various standard conditions, including oxidation by lithium *tert*-butylperoxide²⁸, so an alternative procedure was developed. Deprotonation of the terminal alkyne with lithium bis(trimethylsilyl)amide (LiHMDS) followed by treatment with trimethylborate led to the formation of an alkynylborate intermediate. Addition of *meta*-chloroperbenzoic acid (*m*-CPBA) to the reaction mixture probably formed an intermediate ketene and/or mixed anhydride that was captured by the adjacent alcohol. To the best of our knowledge, this procedure for alkyne oxidation has not yet been reported. Hydrogenolysis of the benzyl esters followed by in situ acidic hydrolysis caused skeletal rearrangement to (–)-**5** in excellent yield.

Introduction of the final, deep C10 oxygen proved challenging. The steric hindrance and ‘bowl’ shape of **5**, in addition to its base-lability, derailed many potential solutions. Compound **5** contains three acidic sites—the hydroxyl, the inner lactone and the outer lactone—but the addition of 3 equivalents of strong base, followed by an acidic quench at –78 °C, caused considerable decomposition and poor mass recovery. We found that treatment with one equivalent of potassium bis(trimethylsilyl)amide (KHMDs) followed by the addition of aqueous 1M HCl enabled full mass recovery; however, it resulted in the clean delivery of the *iso*-bilobalide scaffold (for example, **16a**, Fig. 3) as a result of intramolecular translactonization²⁹. Similarly, we observed trans-lactonization at room temperature when using 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) (bilobalide to *iso*-bilobalide, see Supplementary Information section 3). This facile rearrangement is probably driven

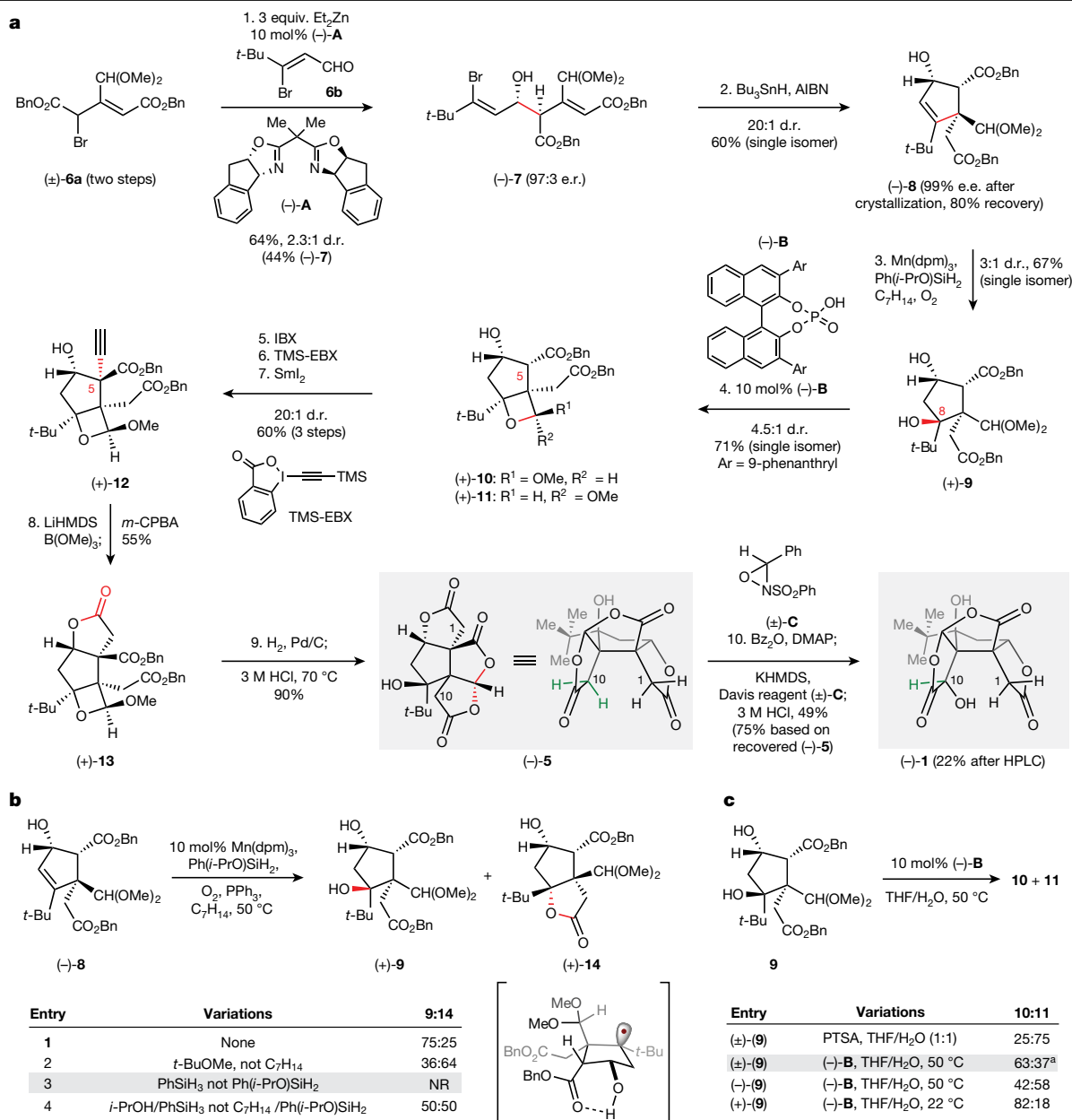


Fig. 2 | Synthesis of (-)-bilobalide. a, Reagents and conditions: (1) **6a**, **6b** (1.2 equiv.), (-)-**A** (10 mol%), Et₂Zn (3.0 equiv.), tetrahydrofuran (THF), -78 °C; (2) Bu₃SnH (1.5 equiv.), azobisisobutyronitrile (AIBN; 0.1 equiv.), toluene, 85 °C; (3) tris(dipivaloylmethanato)manganese (Mn(dpm)₃; 10 mol%), Ph(*i*-PrO)SiH₂ (3.0 equiv.), PPh₃ (1.5 equiv.), methylcyclohexane (C₇H₁₄), O₂ (1 atm), 50 °C; (4) (-)-**B** (10 mol%), THF/H₂O (2:1), 23 °C; (5) IBX (3.0 equiv.), dimethyl sulfoxide, 23 °C; (6) TMS-EBX (3.0 equiv.), TBAF (3.0 equiv.), THF, -78 °C to -20 °C; (7) Sml₂ (8.4

equiv.), THF/H₂O (5:1), 0 °C; (8) LiHMDS (3.0 equiv.), THF, -78 °C; B(OMe)₃ (5.0 equiv.), 23 °C; *m*-CPBA (5.0 equiv.), 0 °C; (9) H₂, Pd/C (10 wt%), MeOH, 23 °C; 3 M HCl (aq.), 80 °C; (10) benzoic anhydride (Bz₂O; 1.5 equiv.), 4-dimethylaminopyridine (DMAP; 1.5 equiv.), THF, 23 °C; KHMDs (3.0 equiv.), -78 °C, (±)-**C** (3 equiv.), -78 °C; 3 M HCl (aq.), 80 °C. **b**, Solvent screen (enabled by Ph(*i*-PrO)SiH₂). **c**, Acid-catalysed oxetane acetal formation. **10**, 39:61 e.r.; **11**, 69:31 e.r.

both by the proximity of the C8 hydroxyl to C4 at the Bürgi–Dunitz angle and by the delocalization of the lactone π system into the adjacent C–O σ^* orbital. Molecular models revealed that the *iso*-bilobalide rearrangement partially folds the skeletal cavity ‘inside out’ (see Fig. 3) to render the inner lactone protons more accessible to reagents. However, neither the alkoxide of **1** nor its *tert*-butyldimethylsilyl (TBS) ether (**16b**) provided substantial bilobalide upon treatment with one equivalent each of base and oxidant (Davis’ oxaziridine, (±)-**17**). Despite the unfolding of the bilobalide cavity and exposure of the C10-proton (highlighted in green), treatment with base still favoured deprotonation of the outer lactone and therefore oxidation provided the isomeric *neo*-bilobalide **15**. We wondered if inner lactone deprotonation required

both increased exposure (rearrangement) and increased acidification. Induction through σ bonds or delocalization of the lactone π system into an adjacent, withdrawn C–O σ^* orbital might acidify the α -protons—that is, stabilize the corresponding enolate³⁰. Conversion of **5** to the isomeric benzoate followed by deprotonation and oxidation at low temperature yielded (-)-**1** with only a trace of **15**. The combination of benzylation and oxidation into one step proved successful, scalable and selective; *neo*-bilobalide (**15**) could not be detected.

The sequence in Fig. 2 has been completed in seven days by one person to produce 0.35 g of (-)-**5**. The same overall strategy disclosed here may be applicable to *G. biloba* congeners including the ginkgolides, some of which are glycine-receptor-selective antagonists⁸. These

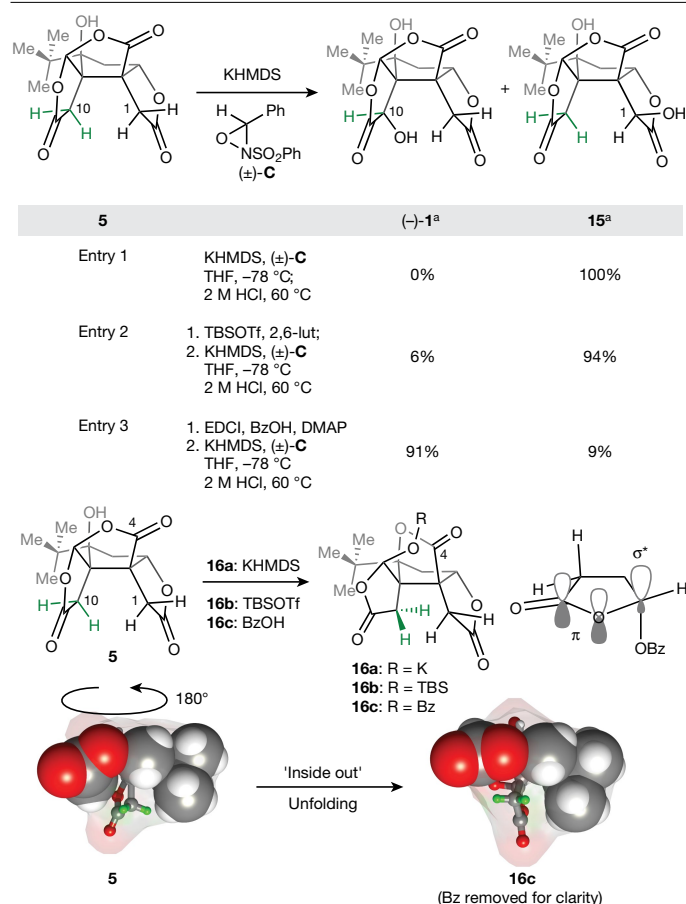


Fig. 3 | Late-stage, regio- and stereoselective oxidation of C10 over C1.

Reagents and conditions for entry 3: (1) 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDCI; 3 equiv.), benzoic acid (BzOH; 3 equiv.), DMAP (0.1 equiv.), THF; (2) KHMDS (1.5 equiv), (±)-C (1.5 equiv); 2 M HCl, 60 °C. ^aBased on conversion.

studies have laid a foundation for new, enabling chemistry—an asymmetric Reformatsky aldol, a solvent-controlled Mukaiyama hydration and a chemoselective alkyne oxidation—and a platform for the functional modification of bilobalide.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1690-5>.

- DeKosky, S. T. et al. *Ginkgo biloba* for prevention of dementia: a randomized controlled trial. *J. Am. Med. Assoc.* **300**, 2253–2262 (2008).
- Wada, K. et al. Studies on the constitution of edible medicinal plants. Isolation and identification of 4-O-methyl-pyridoxine toxic principle from the seed of *Ginkgo biloba*. *Chem. Pharm. Bull.* **36**, 1779–1782 (1988).

- Clarke, T. C., Black, L. I., Stussman, B. J., Barnes, P. M. & Nahin, R. L. *Trends in the Use of Complementary Health Approaches Among Adults: United States, 2002–2012*. National Health Statistics Reports, no. 79 (US Department of Health and Human Services, 2015).
- Kiewert, C. et al. Role of GABAergic antagonism in the neuroprotective effects of bilobalide. *Brain Res.* **1128**, 70–78 (2007).
- Fernandez, F. et al. Pharmacotherapy for cognitive impairment in a mouse model of Down syndrome. *Nat. Neurosci.* **10**, 411–413 (2007).
- van Beek, T. A. & Taylor, L. T. Sample preparation of standardized extracts of *Ginkgo biloba* by supercritical fluid extraction. *Phytochem. Anal.* **7**, 185–191 (1996).
- Lynch, J. W. & Chen, X. Subunit-specific potentiation of recombinant glycine receptors by NV-31, a bilobalide-derived compound. *Neurosci. Lett.* **435**, 147–151 (2008).
- Ivic, L. et al. Terpene trilactones from *Ginkgo biloba* are antagonists of cortical glycine and GABA_A receptors. *J. Biol. Chem.* **278**, 49279–49285 (2003).
- Huang, S. H. et al. Bilobalide, a sesquiterpene trilactone from *Ginkgo biloba*, is an antagonist at recombinant $\alpha_1\beta_2\gamma_2$. *Eur. J. Pharm.* **464**, 1–8 (2003).
- Thompson, A. J., McGonigle, I., Duke, R., Johnston, G. A. R. & Lummis, S. C. R. A single amino acid determines the toxicity of *Ginkgo biloba* extracts. *FASEB J.* **26**, 1884–1891 (2012).
- Nakanishi, K. et al. Structure of bilobalide, a rare *tert*-butyl containing sesquiterpenoid related to the C₂₀-ginkgolides. *J. Am. Chem. Soc.* **93**, 3544–3546 (1971).
- Strømgaard, K. & Nakanishi, K. Chemistry and biology of terpene trilactones from *Ginkgo biloba*. *Angew. Chem. Int. Ed.* **43**, 1640–1658 (2004).
- Vale, S. Subarachnoid haemorrhage associated with *Ginkgo biloba*. *Lancet* **352**, 36 (1998).
- Ng, C. C., Duke, R. K., Hinton, T. & Johnston, G. A. R. Effects of bilobalide, ginkgolide B and picrotoxinin on GABA_A receptor modulation by structurally diverse positive modulators. *Eur. J. Pharm.* **806**, 83–90 (2017).
- Weinges, K., Hepp, M., Huber-Patz, U., Rodewald, H. & Irngartinger, H. Chemistry of ginkgolides. 1. 10-acetyl-1-methoxycarbonyl-2,3,14,15,16-pentanorginkgolide-A, an intermediate for the synthesis of bilobalide. *Liebigs Ann. Chem.* 1057–1066 (1986).
- Harrison, T., Myers, P. L. & Pattenden, G. Radical cyclisations onto 2(5H)-furanone and maleate electrophore. An approach to the spiro- and linear-fused γ -lactone ring systems found in the ginkgolides. *Tetrahedron* **45**, 5247–5262 (1989).
- Emsermann, J. & Opatz, T. Photochemical approaches to the bilobalide core. *Eur. J. Org. Chem.* 3362–3372 (2017).
- Corey, E. J. & Su, W. G. Total synthesis of a C₁₅ ginkgolide, (±)-bilobalide. *J. Am. Chem. Soc.* **109**, 7534–7536 (1987).
- Corey, E. J. & Su, W. G. Enantioselective total synthesis of Bilobalide, a C₁₅ ginkgolide. *Tetrahedron Lett.* **29**, 3423–3426 (1988).
- Crimmins, M. T., Jung, D. K. & Gray, J. L. Synthetic studies on the ginkgolides: total synthesis of (±)-bilobalide. *J. Am. Chem. Soc.* **115**, 3146–3155 (1993).
- Fernández-Ibáñez, M. Á., Maciá, B., Alonso, D. A. & Pastor, I. M. Recent advances in the catalytic enantioselective Reformatsky reaction. *Eur. J. Org. Chem.* 7028–7034 (2013).
- Wolf, C. & Moskowit, M. Bisoxazolidine-catalyzed enantioselective Reformatsky reaction. *J. Org. Chem.* **76**, 6372–6376 (2011).
- Crossley, S. W. M., Obradors, C., Martínez, R. M. & Shenvi, R. A. Mn-, Fe-, and Co-catalyzed radical hydrofunctionalizations of olefins. *Chem. Rev.* **116**, 8912–9000 (2016).
- Obradors, C., Martínez, R. M. & Shenvi, R. A. Ph. (i-PrO)SiH₂: a remarkable reductant for metal-catalyzed hydrogen atom transfers. *J. Am. Chem. Soc.* **138**, 4962–4971 (2016).
- Maier, G., Pfiem, S., Schäfer, U. & Matusch, R. Tetra-*tert*-butyltetrahydrene. *Angew. Chem. Int. Edn Engl.* **17**, 520–521 (1978).
- Fernández-González, D. F., Brand, J. P. & Waser, J. Ethynyl-1,2-benziodoxol-3(1H)-one (EBX): an exceptional reagent for the ethynylation of keto, cyano, and nitro esters. *Chem. Eur. J.* **16**, 9457–9461 (2010).
- Keck, G. E. & Wagner, C. A. The first directed reduction of β -alkoxy ketones to *anti*-1,3-diol monoethers: identification of spectator and director alkoxy group. *Org. Lett.* **2**, 2307–2309 (2000).
- Julia, M., Saint-Jalmes, V. P. & Verpeaux, J. N. Oxidation of carbanions with lithium *tert*-butyl peroxide. *Synlett* **1993**, 233–234 (1993).
- Weinges, K., Hepp, M., Huber-Patz, U. & Irngartinger, H. Chemistry of ginkgolides. III. Bilobalide/isobilobalide. Structure determination by X-ray analysis. *Liebigs Ann. Chem.* 1079–1085 (1986).
- Byun, K., Mo, Y. & Gao, J. New insight on the origin of the unusual acidity of Meldrum's Acid from ab initio and combined QM/MM simulation study. *J. Am. Chem. Soc.* **123**, 3974–3979 (2001).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Data availability

All data is available in the text of this Article or its Supplementary Information. Structural parameters are available from the Cambridge Crystallographic Data Centre (CCDC) under the following reference numbers: (–)-**5**, CCDC 1911131; (–)-**8**, CCDC 1911128; **12**, CCDC 1911129; and **16c**, CCDC 1911127.

Acknowledgements We thank P. Baran and K. Engle for conversations, and the Engle laboratory for donations of chiral phosphoric acids, including (–)-**B**. A. Rheingold, C. Moore and M. Gembicky are acknowledged for X-ray crystallographic analysis. We thank J. Chen and B. Sanchez in the Scripps Research Automated Synthesis Facility for purification assistance and for analysis of chiral non-racemic compounds. Support was provided by the National Institutes of Health (R35 GM122606) and the Uehara Memorial Foundation; additional support was provided by Eli Lilly, Novartis, Bristol-Myers Squibb, Amgen, Boehringer-Ingelheim, the Sloan Foundation and the Baxter Foundation.

Author contributions R.A.S., M.A.B., R.M.D. and M.O. conceived the project. R.A.S. directed the research, and R.A.S., M.O., M.A.B. and R.M.D. composed the manuscript and the Supporting Information section. M.O., M.A.B. and R.M.D. completed a first-generation synthesis of *rac*-**1**. M.A.B. conceived and developed the catalytic asymmetric synthesis of (–)-**7**. M.A.B. observed, designed and optimized the parallel kinetic resolution of *rac*-**9**. M.O. and R.M.D. screened and optimized conditions for the alkyne oxidation of *rac*-**12** and (+)-**12**. R.M.D. developed the hydration of *rac*-**8** and (–)-**8** and optimized scale-up campaigns of *rac*-**5** and (–)-**5**. M.A.B. and R.M.D. conducted large-scale syntheses of *rac*-**5** and (–)-**5**. M.A.B. and R.M.D. investigated the rearrangement of *rac*-**5** and (–)-**5** to **16a–c**. M.O. discovered an oxidation of *rac*-**5** to *rac*-**1**. M.A.B. investigated the rearrangement of *rac*-**5** and (–)-**5** to **16b** and **16c**, and discovered conditions that were utilized for the oxidation of *rac*-**5** and (–)-**5** to *rac*-**1** and (–)-**1**; M.A.B. and R.M.D. both optimized this process.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1690-5>.

Correspondence and requests for materials should be addressed to M.O. or R.A.S.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Twofold expansion of the Indo-Pacific warm pool warps the MJO life cycle

<https://doi.org/10.1038/s41586-019-1764-4>

Received: 21 May 2019

Accepted: 17 September 2019

Published online: 27 November 2019

M. K. Roxy^{1,2*}, Panini Dasgupta^{1,3}, Michael J. McPhaden², Tamaki Suematsu⁴, Chidong Zhang² & Daehyun Kim⁵

The Madden–Julian Oscillation (MJO) is the most dominant mode of subseasonal variability in the tropics, characterized by an eastward-moving band of rain clouds. The MJO modulates the El Niño Southern Oscillation¹, tropical cyclones^{2,3} and the monsoons^{4–10}, and contributes to severe weather events over Asia, Australia, Africa, Europe and the Americas. MJO events travel a distance of 12,000–20,000 km across the tropical oceans, covering a region that has been warming during the twentieth and early twenty-first centuries in response to increased anthropogenic emissions of greenhouse gases¹¹, and is projected to warm further. However, the impact of this warming on the MJO life cycle is largely unknown. Here we show that rapid warming over the tropical oceans during 1981–2018 has warped the MJO life cycle, with its residence time decreasing over the Indian Ocean by 3–4 days, and increasing over the Indo-Pacific Maritime Continent by 5–6 days. We find that these changes in the MJO life cycle are associated with a twofold expansion of the Indo-Pacific warm pool, the largest expanse of the warmest ocean temperatures on Earth. The warm pool has been expanding on average by $2.3 \times 10^5 \text{ km}^2$ (the size of Washington State) per year during 1900–2018 and at an accelerated average rate of $4 \times 10^5 \text{ km}^2$ (the size of California) per year during 1981–2018. The changes in the Indo-Pacific warm pool and the MJO are related to increased rainfall over southeast Asia, northern Australia, Southwest Africa and the Amazon, and drying over the west coast of the United States and Ecuador.

Each year, weather variability at subseasonal to seasonal timescales costs the global economy over US\$2 trillion, with US\$700 billion alone in the United States (3.4% of US GDP in 2018)^{12,13}. The MJO contributes to more than 55% of this weather variability over the tropics¹⁴, and modulates the Asian^{4,5}, Australian⁶, African⁷ and American monsoons^{8–10}, tropical cyclogenesis^{2,3} and the El Niño Southern Oscillation (ENSO)¹. The phase and strength of the MJO at a given location can enhance or suppress tropical rainfall variability, modulating or triggering extreme weather events including hurricanes, droughts, flooding, heat waves and cold surges¹⁵. The MJO can also lead to marked effects at mid-latitudes, and is a strong contributor to extreme events in the United States and Europe^{16,17}. The intensity and propagation of the MJO is shown to influence the circulation pattern in the Arctic stratosphere and the polar vortex¹⁸, emphasizing the far-reaching effect of the MJO on the Earth's climate system.

The MJO is an ocean–atmosphere-coupled phenomenon, characterized by eastward moving disturbances of clouds, rainfall, winds and pressure along the Equator. It is the most dominant mode of subseasonal variability in the tropics¹⁹. Using observations and model simulations, previous studies have attempted to understand changes in the MJO in a warming climate²⁰. A link was found between increasing carbon emission and changes in the intensity, frequency and propagation of the MJO over the last few decades of the twentieth century^{20,21}, although

there is considerable uncertainty as to the extent of the changes and the mechanisms involved. A statistical reconstruction of MJO activity over 1905–2008 using tropical surface pressures shows a 13% increase per century in MJO amplitude²². The reconstructed MJO activity agrees with satellite-observed (since 1979) MJO variability on decadal timescales, but the trends disagree after 1997²³, which adds to the considerable uncertainty in these long-term trends. Studies also suggest an increasing trend in MJO frequency after the mid-1970s^{24,25}, which has been linked to long-term warming in the tropical oceans²⁶. Numerical model experiments under idealized global warming scenarios indicate that increasing the surface temperature over the tropical oceans results in an organized MJO activity with a faster eastward propagation^{21,27}, although an understanding based on observations is pending.

Typically, the MJO events are initiated over the Indian Ocean and move eastward over the Maritime Continent to the Pacific (Extended Data Fig. 1). Some of these events weaken or break down over the Maritime Continent or the central Pacific²⁸, but others propagate further to the east Pacific and occasionally continue into the Atlantic²⁹. On average, MJO events travel a zonal distance of 12,000–20,000 km (7,500–12,500 miles) over the generally warm tropical oceans. This entire stretch of the tropical ocean has been warming during the twentieth and early twenty-first centuries in response to greenhouse gas forcing¹¹, and is projected to warm further in the future. The rapid warming across

¹Centre for Climate Change Research, Indian Institute of Tropical Meteorology, Pune, India. ²Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, Seattle, WA, USA. ³Department of Meteorology and Oceanography, College of Science and Technology, Andhra University, Visakhapatnam, Andhra Pradesh, India. ⁴Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Chiba, Japan. ⁵Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA. *e-mail: roxy@tropmet.res.in

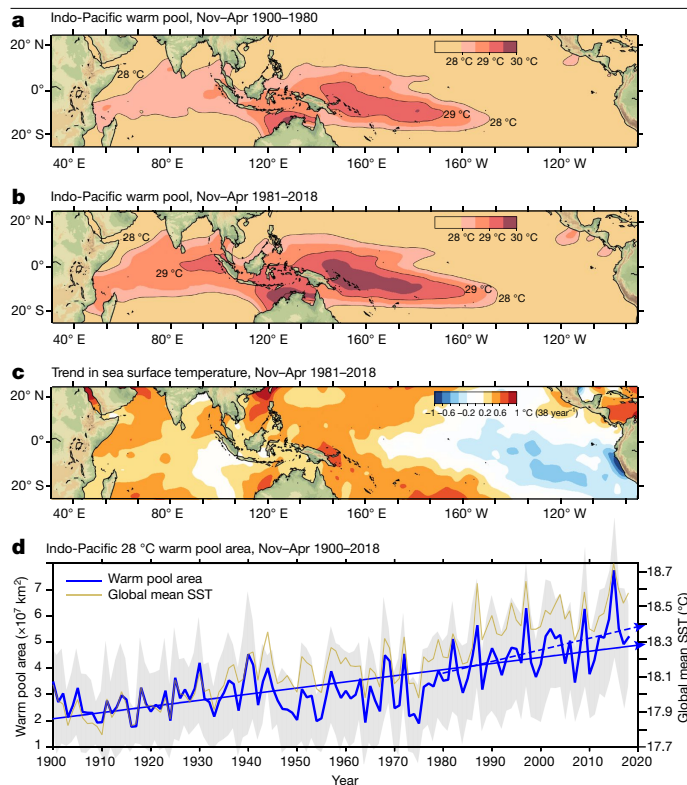


Fig. 1 | A twofold expansion of the warm pool. **a, b,** Indo-Pacific warm pool with its characteristic permanently warm SSTs of greater than 28 °C for the period 1900–1980 (**a**) and 1981–2018 (**b**). The observed warm pool expansion is almost twofold, from an area of 2.2×10^7 km² during 1900–1980, to an area of 4×10^7 km² during 1981–2018. The warm pool area is estimated as the surface area covered by climatological 28 °C isotherm of SST, during November–April, in the tropical Indo-Pacific region within 40° E to 140° W, 25° S to 25° N. **c,** The observed trend in SST (°C per 38 years) during 1981–2018, for November–April. **d,** Time series of the warm pool area during 1900–2018. Theil–Sen trend estimates are overlaid on the time series for the entire period (solid blue line, Sen slope of 2.25×10^5 km² per year) and for 1981–2018 (dashed blue line, Sen slope of 4.14×10^5 km² per year). The positive trend in warm pool area is significant at the 95% confidence level, according to the Mann–Kendall test. The grey shade overlaid on the time series represents \pm two standard deviations of the warm pool area, based on monthly SST values. The yellow line represents global mean SSTs (°C) averaged for November–April. Warm pool SST values and area are based on HadISST dataset. The Generic Mapping Tools (GMT, <https://github.com/GenericMappingTools/gmt>) was used to create the topographic map, with the topography data from ETOPO1 Global Relief Model (<https://www.ngdc.noaa.gov/mgg/global/global.html>).

the tropical basins is not uniform. In the equatorial belt, the largest warming during November–April when the MJO is active is observed over the Indo-Pacific warm pool (Fig. 1). This warm pool is the largest region of permanently warm sea surface temperatures (SSTs >28 °C), covering an area greater than 2.7×10^7 km² (see Methods), over which there is vigorous deep convection. The tropical ocean warming has led to an expansion of this warm pool, particularly in recent decades. Even though there is a preliminary understanding of the general changes in MJO amplitude and frequency in a warming climate, it is not known how the non-uniform ocean warming associated with the expanding warm pool may affect the MJO regionally.

In our study, we find a twofold expansion of the Indo-Pacific warm pool during 1981–2018, in comparison to 1900–1980, with the largest warming occurring over the western Pacific. We show that this warm pool expansion has led to significant changes in the life cycle of MJO events over the Indo-Pacific region. Although the total period of the MJO does not show any detectable trends, its residence time

(MJO phase duration) over the Indian Ocean has been reduced by 3–4 days while over the Maritime Continent its residence time has increased by 5–6 days. Essentially, this means that MJO-related convective activity has grown shorter over the Indian Ocean while the convection over the Maritime Continent is being prolonged.

Observed changes in MJO life cycle

We select the MJO events that exhibit strong coupling between tropical convection and large-scale circulation; prominent active eastward propagation; and an amplitude of the real-time multivariate (RMM) MJO index³⁰ that is greater than one for November–April, 1981–2018 (see Methods). From its normal initiation in the Indian Ocean (RMM phase 1), the MJO propagates into the central Pacific and beyond (phase 8) in about 30–60 days (Extended Data Fig. 2). We compute the average number of days of the selected MJO in each RMM phase to describe the MJO phase duration over the tropical ocean basins. In the RMM index, interannual variations, including those associated with ENSO³⁰, have been removed. This makes it suitable for our investigation focusing on the changes in the MJO related to global warming.

Figure 2 shows the time series of the MJO phase duration and how it has changed over time. The average period of the MJO does not exhibit any detectable trend and broadly remains within the normal timescale of 30–60 days (Extended Data Fig. 2). However, closer inspection (Fig. 2) shows significant changes in individual phases, which essentially are offset while averaging over the entire MJO domain. Over the Indian Ocean (RMM phases 1, 2 and 3), the MJO phase duration decreases by 3–4 days, from an average of 19 days (during 1981–1999) to 15.4 days (during 2000–2018) (Fig. 2a, b). Over the Maritime Continent and the west Pacific (RMM phases 5, 6 and 7), the MJO phase duration increases by 5–6 days, from an average of 17.5 days to 23 days (Fig. 2c, d). The observed trends are statistically significant at the 95% confidence level. The changes are consistent with those documented by previous studies that compared the MJO activity across different RMM phases, using observations and climate model experiments^{31,32}. This means that during recent decades, convective cloud bands associated with the MJO linger over the Indian Ocean for a shorter period, while they persist longer over the Maritime Continent and the west Pacific.

The role of Indo-Pacific warming

SST variations mediate the exchange of heat across the air–sea interface. High SSTs over the tropics are usually accompanied by enhanced convective activity³³. Being an ocean–atmosphere-coupled convective phenomenon, MJO activity is therefore highly dependent on tropical SSTs, with higher MJO activity typically occurring when SSTs are higher²⁶. Previous studies have shown accelerated warming over the Indo-Pacific warm pool and its expansion^{11,34}, which can potentially have an impact on the MJO characteristics. To examine the changes in the warm pool, we estimated the surface area covered by the climatological 28 °C isotherm of SST, during November–April (Fig. 1), in the tropical Indo-Pacific region within 40° E to 140° W, 25° S to 25° N. We show that tropical SST warming has led to an almost twofold expansion of the Indo-Pacific warm pool, from an area of 2.2×10^7 km² during 1900–1980, to an area of 4×10^7 km² during 1981–2018 (Fig. 1a, b, d). The warm pool expansion is non-uniform, with the SST warming more pronounced over the west Pacific in contrast to the Indian Ocean (Fig. 1c). The difference in warm pool expansion trends between the 1900–1980 and 1981–2018 periods is statistically significant at the 95% confidence levels. The shift in warm pool SSTs during the 1977–1980 period co-occur with the shift in global mean SSTs at the same time (Fig. 1d), followed by an accelerated surface warming as a response to anthropogenic emissions³⁵. It is important to note that the shift in SSTs also coincides with the positive phase of the Pacific Decadal Oscillation (PDO). A comparison of the warm pool area using

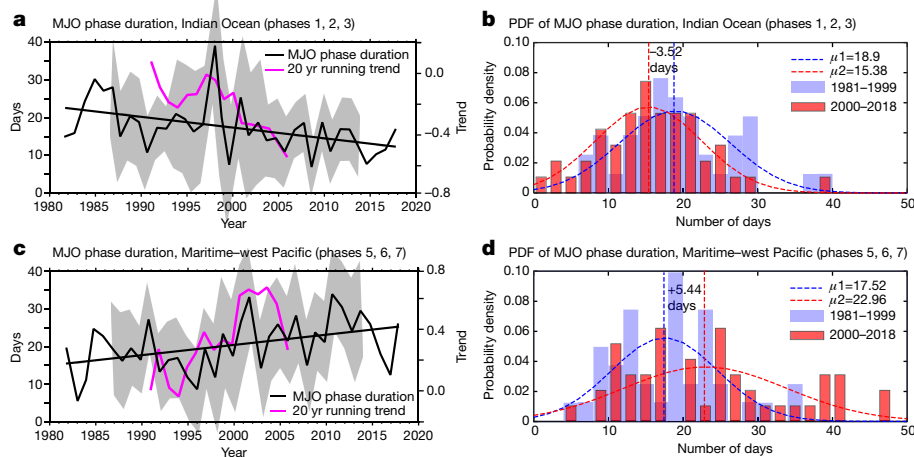


Fig. 2 | Changes in the MJO life cycle. a–d, Time series (black line) and distribution of average yearly phase duration (in days) of MJO events during 1981–2018 over the Indian Ocean (**a**, **b**; RMM phases 1, 2 and 3) and the Maritime–west Pacific region (**c**, **d**; RMM phases 5, 6 and 7). Grey shading in **a**, **c** is \pm two standard deviations of the MJO phase duration over a 10-year moving window. Pink lines overlaid on the time series represent the 20-year running trend of MJO phase duration (in days per year). Mann–Kendall test for the time series indicates that the trends are significant at the 95% confidence level. The

phase duration distribution compares the probability density function (PDF) of MJO phase duration during the earlier period (1981–1999) and later period (2000–2018), where μ_1 and μ_2 represent the mean number of days. $\mu_2 - \mu_1$ indicates the change in MJO phase duration, with a decrease of 3–4 days over the Indian Ocean and increase of 5–6 days over the Maritime–west Pacific region. A Mann–Whitney *U*-test on the difference in the phase duration distributions in **b**, **d** shows that the difference is statistically robust ($P < 0.05$), implying that the null hypothesis can be rejected.

multiple SST datasets shows that the changes in warm pool area presented here are robust (Extended Data Fig. 3a). A breakpoint analysis confirms that the shifts to higher warm pool values occurred during 1979–1980 (Extended Data Fig. 3b, c).

The changes in the MJO phase duration (phases 5, 6 and 7) appear to be significantly correlated (Fig. 3 and Extended Data Fig. 4) to the changes in SST collocated over the west Pacific warm pool, where the warming trends and the background mean SSTs are the largest. The fact that the correlation is significant even after the trends are removed suggests that the mechanisms working on interannual and longer timescales are similar. Although SST warming is observed in the Indian Ocean also, it is interesting that these SST trends do not show any significant correlation with the MJO phase duration (phases 5, 6 and 7). This might mean that the observed changes in the MJO phase duration are driven by SST changes in the west Pacific. In fact, an investigation of the atmospheric circulation shows enhanced convective activity and a strengthening of low-level westerlies over the west Pacific (120° E to 160° E) associated with trends in the MJO phase duration (Fig. 3b). The enhanced convective activity over the west Pacific is compensated by subsidence over the central and west Indian Ocean (40° E to 70° E). Pohl and Matthews²⁵ hypothesize that on interannual timescales when the west Pacific is warmer than normal, the latent heat release over the moist convective region decreases the effective static stability of the atmosphere and slows down the MJO over the warm pool. The long-term changes in the MJO phase duration and associated ocean–atmospheric interactions discerned here are consistent with the physical mechanisms observed for the MJO phase duration on interannual timescales^{25,36}.

MJO variability and propagation are largely linked to moist static energy in the atmospheric column^{36–38}. For a detailed examination of factors driving the observed trends in the MJO phase duration, we inspected the specific humidity and temperature profiles independently (Fig. 3c). Whereas the MJO trends (phases 5, 6 and 7) exhibit a positive correlation with tropospheric temperatures over the warm pool from 90° E to 170° E, the specific humidity anomalies show a significantly negative correlation over the Indian Ocean and positive correlation over the west Pacific. This indicates that while the warm SST trend in the west Pacific prolongs the local convective activity, it also drives dry air subsidence over the Indian Ocean

(along with the moisture advected away from the basin), shortening the residence time of the MJO over that region. Hence, although the SST over the entire Indo-Pacific is warming, it appears that the MJO response is more sensitive to the west Pacific SST, possibly because the SST trends and background mean values are relatively larger over this region during November–April. Meanwhile, the low-level winds associated with the observed changes in phase duration are westerly over the Indian Ocean (Fig. 3b), converging into the west Pacific. This indicates that the prolonged residence time of the MJO over the Maritime Continent may be supported by moisture supply from both local (west Pacific) and remote (Indian Ocean) sources. Extended Data Figure 5 shows a significant increase in tropospheric moisture (900–400 hPa levels) over the Maritime Continent–west Pacific warm pool region and a reduction in moisture over the Indian Ocean. This is consistent with previous studies³⁹ that suggest that the moisture gradient in the lower troposphere over the Indo-Pacific warm pool assists the eastward propagation of the MJO.

A comparison of the MJO phase duration over the Maritime Continent and west Pacific warm pool area (120° E to 160° E, 25° S to 25° N, highlighted region in Fig. 3, phases 5, 6 and 7) demonstrates a considerable correlation (Pearson correlation, $r = 0.42$; Kendall rank correlation, $\tau = 0.3$) statistically significant at the 95% confidence level (Fig. 3d). The MJO phase duration over the Indian Ocean (phases 1, 2 and 3) also shows a significant negative correlation with the west Pacific ($r = -0.33$), suggesting that the MJO changes over the Indian Ocean are also largely driven by SST warming over the west Pacific. A correlation with the trends removed from both the time series still shows statistical significance at the 90% confidence level, and it can be argued that the results of this analysis strongly hold, even if the large values of the correlation coefficient are due to the existence of a real trend. Furthermore, the mean surface temperatures over the west Pacific also exhibit an inter-annual variability and long-term change similar to that of the warm pool expansion (Fig. 3d, $r = 0.97$, $\tau = 0.86$). The results presented here establish a clear role of warm pool expansion and increasing SSTs in shortening the residence time of the MJO over the Indian Ocean by 3–4 days and prolonging it over the Maritime Continent by 5–6 days (Extended Data Fig. 6). Such a large change in the MJO phase duration may have direct implications on the global weather and climate, which are tightly linked to these MJO phases.

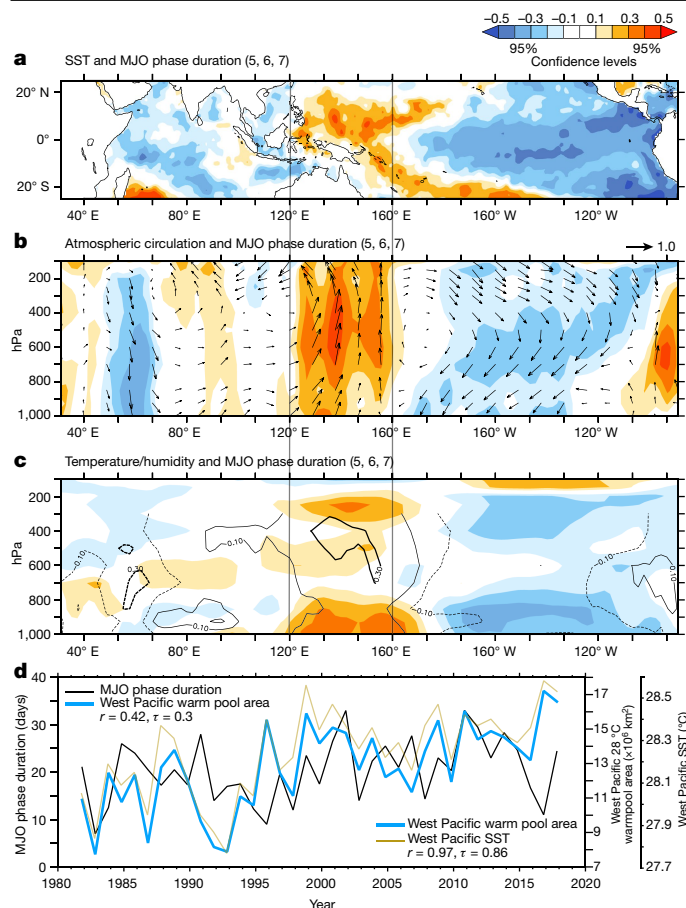


Fig. 3 | Correlation between MJO phase duration and ocean-atmosphere conditions. **a–c**, Correlation between the MJO phase duration (phases 5, 6 and 7) and SST anomalies (**a**), winds and vertical velocity (**b**), and air temperature (colours) and specific humidity (contours) (**c**) at each grid point over the Indo-Pacific basin for November–April, during 1981–2018 ($n = 37$). The correlation analysis is performed after removing the trend and the ENSO variability from the time series. Colour shading denotes correlation coefficients, with the significance at the 95% confidence level noted below the colour scale (above **a**). Vector arrow lengths are proportional to correlation coefficient according to the scale on top of **b**. Thick contours in **c** denote correlation coefficients significant at the 95% confidence level. The region within the solid black lines highlights the west Pacific warm pool region (120°E to 160°E) where the ocean–atmospheric changes related to the MJO phase duration are the largest, and consistent across the various parameters. The longitude–pressure plots are averaged over 10°S to 10°N. **d**, Time series of MJO phase duration (phases 5, 6 and 7) and the surface area (km²) enclosed by the 28 °C isotherm of SST over the west Pacific (120°E to 160°E, 25°S to 25°N), during November–April, 1981–2018. The Kendall rank correlation test (two tailed) for the two variables provided a tau coefficient (τ) of 0.3. The Kendall (τ) and Pearson (r) correlation coefficients shown are significant at the 95% confidence level (significant at the 90% confidence level after removing the trends, $n = 37$). The yellow line overlaid on the time series represents the yearly mean SST over the west Pacific. The Kendall rank correlation test for the west Pacific warm pool area and SST provided a tau coefficient (τ) of 0.86, significant at the 95% confidence level. PyFerret (<http://ferret.pmel.noaa.gov/Ferret/>) was used to generate the map and the plots.

Impacts on global climate

To assess the potential impacts of the observed changes in the MJO phase duration on global climate, we performed a correlation analysis with the rainfall anomalies at each location across the globe, after removing the trends and ENSO-related variability. Figure 4a shows significantly large correlation between observed changes in the MJO phase duration and rainfall variability over tropical and mid-latitude

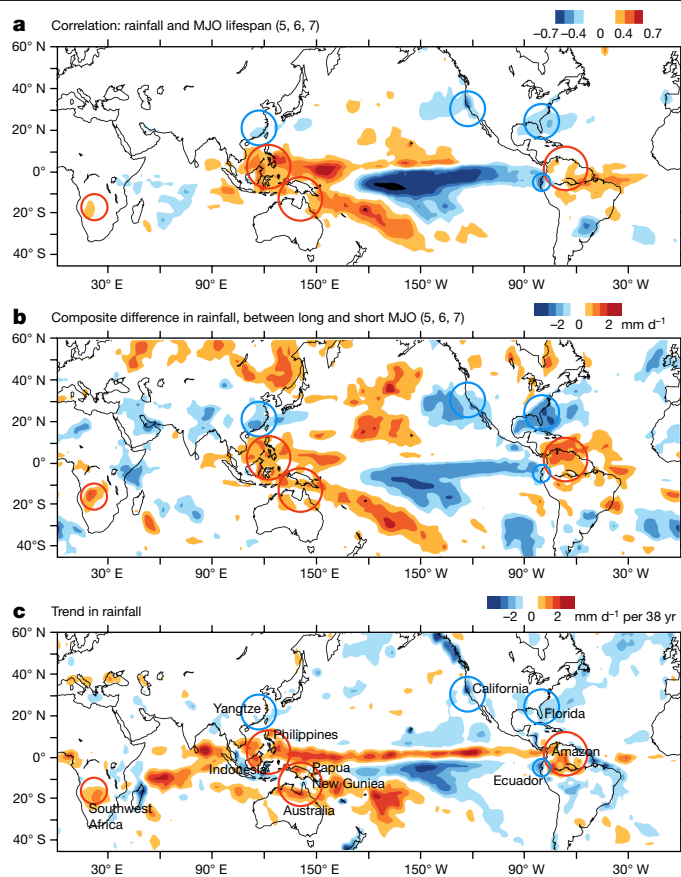


Fig. 4 | Changes in global rainfall in response to the changes in MJO phase duration. **a**, Correlation between the MJO (phases 5, 6 and 7) phase duration and rainfall anomalies for November–April, during 1981–2018. **b**, Composite difference between years when the MJO phase duration (phases 5, 6 and 7) is long and short (above and below one standard deviation). **c**, Observed trend in rainfall (mm day⁻¹ per 38 years) during the same period. The correlation and composite analyses are performed after removing the trend and the ENSO variability from the time series. Colour shading denotes correlation coefficients and trends significant at the 95% confidence level. The circled regions indicate large continental areas where the trends in rainfall are consistent with the correlation and composite analyses. Red circles indicate increasing rainfall and blue circles indicate decreasing rainfall associated with the observed changes in the MJO phase duration. Rainfall values are based on the GPCP dataset. The Generic Mapping Tools (GMT, <https://github.com/GenericMappingTools/gmt>) was used to create the map.

regions. The changes in the MJO phase duration over the Indo-Pacific are associated with enhanced rainfall over the Maritime Continent–west Pacific region, the Amazon basin in South America, southwest Africa and northern Australia (colour shades in Fig. 4a indicate correlation coefficients significant at the 95% confidence level). Meanwhile, the changes in MJO phase duration indicate a strong link with reduced rainfall over central and east Pacific, east Africa, the Ganges basin in India, Yangtze basin in China, and the east and west coasts of the United States.

Notably, a trend analysis of rainfall for November–April shows consistent changes over some of these regions (Fig. 4c). An increase in mean rainfall is observed over most of the Maritime Continent, including southeast Asia (Indonesia, Philippines and Papua New Guinea), northern Australia, west Pacific, Amazon basin and southwest Africa. A decline in rainfall is observed over the central Pacific, Ecuador and along the west coast of the United States (California). A slight decrease in rainfall is observed over the Yangtze basin in China and east coast of the United States (Florida), consistent with changes in the MJO phase duration. The observed impacts on rainfall are consistent with the MJO

impacts on interannual timescales reported by previous studies¹⁵, which means that similar processes are operating at interannual and lower frequency timescales (Extended Data Fig. 7). We confirm this with a composite analysis of the MJO events with longer phase duration for phases 5, 6 and 7 (standard deviation greater than one), which shows similar results as for the correlation analysis and the trends (Fig. 4b).

The recent California droughts (2013–2014, during which the MJO was in phases 5, 6 and 7 for 25–28 days), southeast Asia floods (in 2011, during which the MJO was in phases 5, 6 and 7 for 30 days) and east Africa droughts (2011) occurred during those years when the MJO phase duration was longer over the Maritime Continent and the west Pacific (Fig. 2c). Extreme flooding events in Brazil, such as the 2011 Rio de Janeiro floods, have been linked to a strong MJO interacting with the South Atlantic Convergence Zone¹⁰. It cannot be ruled out that the same mean state change (namely warm pool expansion) can affect both the MJO and the regional rainfall changes presented here. In addition, large-scale changes in circulation due to Indo-Pacific warming⁴⁰ and the phase of the PDO could also interact with the MJO to influence the regional rainfall changes observed here. Regardless of their inter-relationship, we can certainly say that the Indo-Pacific warm pool expansion is not only changing the MJO but also these regional precipitation anomalies, either synergistically through the MJO or through independent pathways. Although we have not investigated the dynamics behind these events individually, we cannot overemphasize the need to closely monitor the changes in the Indo-Pacific warm pool for triggering or intensifying severe weather events in the future. Maintaining and enhancing existing ocean observational arrays over the Indian and Pacific basins and extending it to the straits in the southeast Asian maritime region is hence a high priority^{41,42}. Climate model projections suggest further warming of the warm pool region, which may intensify the observed changes in the MJO life cycle in future. However, state-of-the-art climate models fail to accurately simulate the observed distribution of SST changes over the Indo-Pacific even in the present climate, and hence may need further improvement (for example, via the subseasonal to seasonal prediction project^{42,43}) in order to meet the challenges presented by a warming world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1764-4>.

- McPhaden, M. J. Genesis and evolution of the 1997–98 El Niño. *Science* **283**, 950–954 (1999).
- Maloney, E. D. & Hartmann, D. L. Modulation of eastern North Pacific hurricanes by the Madden–Julian oscillation. *J. Clim.* **13**, 1451–1460 (2000).
- Klotzbach, P. J. & Oliver, E. C. Modulation of Atlantic basin tropical cyclone activity by the Madden–Julian oscillation (MJO) from 1905 to 2011. *J. Clim.* **28**, 204–217 (2015).
- Joseph, S., Sahai, A. & Goswami, B. Eastward propagating MJO during boreal summer and Indian monsoon droughts. *Clim. Dyn.* **32**, 1139–1153 (2009).
- Jia, X., Chen, L., Ren, F. & Li, C. Impacts of the MJO on winter rainfall and circulation in China. *Adv. Atmos. Sci.* **28**, 521–533 (2011).
- Wheeler, M. C., Hendon, H. H., Cleland, S., Meinke, H. & Donald, A. Impacts of the Madden–Julian oscillation on Australian rainfall and circulation. *J. Clim.* **22**, 1482–1498 (2009).
- Pohl, B. & Camberlin, P. Influence of the Madden–Julian oscillation on East African rainfall. I: intraseasonal variability and regional dependency. *Q. J. R. Meteorol. Soc.* **132**, 2521–2539 (2006).
- Lorenz, D. J. & Hartmann, D. L. The effect of the MJO on the North American monsoon. *J. Clim.* **19**, 333–343 (2006).
- Grimm, A. M. Madden–Julian Oscillation impacts on South American summer monsoon season: precipitation anomalies, extreme events, teleconnections, and role in the MJO cycle. *Clim. Dyn.* **53**, 1–26 (2019).

- Carvalho, L. M. V., Jones, C. & Liebmann, B. The South Atlantic convergence zone: intensity, form, persistence, and relationships with intraseasonal to interannual activity and extreme rainfall. *J. Clim.* **17**, 88–108 (2004).
- Weller, E. et al. Human-caused Indo-Pacific warm pool expansion. *Sci. Adv.* **2**, e1501719 (2016).
- Lazo, J. K., Lawson, M., Larsen, P. H. & Waldman, D. M. US economic sensitivity to weather variability. *Bull. Am. Meteorol. Soc.* **92**, 709–720 (2011).
- Bertrand, J.-L. & Brusset, X. Managing the financial consequences of weather variability. *J. Asset Manag.* **19**, 301–315 (2018).
- Kessler, W. S. EOF representations of the Madden–Julian oscillation and its connection with ENSO. *J. Clim.* **14**, 3055–3061 (2001).
- Zhang, C. Madden–Julian oscillation: bridging weather and climate. *Bull. Am. Meteorol. Soc.* **94**, 1849–1870 (2013).
- Cassou, C. Intraseasonal interaction between the Madden–Julian Oscillation and the North Atlantic Oscillation. *Nature* **455**, 523–527 (2008).
- Stan, C. et al. Review of tropical-extratropical teleconnections on intraseasonal time scales. *Rev. Geophys.* **55**, 902–937 (2017).
- Garfinkel, C. I., Feldstein, S. B., Waugh, D. W., Yoo, C. & Lee, S. Observed connection between stratospheric sudden warmings and the Madden–Julian Oscillation. *Geophys. Res. Lett.* **39**, L18807 (2012).
- Madden, R. A. & Julian, P. R. Observations of the 40–50-day tropical oscillation—a review. *Mon. Weath. Rev.* **122**, 814–837 (1994).
- Maloney, E. D., Adames, Á. F. & Bui, H. X. Madden–Julian oscillation changes under anthropogenic warming. *Nat. Clim. Change* **9**, 26–33 (2019).
- Adames, Á. F., Kim, D., Sobel, A. H., Del Genio, A. & Wu, J. Changes in the structure and propagation of the MJO with increasing CO₂. *J. Adv. Model. Earth Syst.* **9**, 1251–1268 (2017).
- Oliver, E. C. & Thompson, K. R. A reconstruction of Madden–Julian Oscillation variability from 1905 to 2008. *J. Clim.* **25**, 1996–2019 (2012).
- Oliver, E. C. Blind use of reanalysis data: apparent trends in Madden–Julian Oscillation activity driven by observational changes. *Int. J. Climatol.* **36**, 3458–3468 (2016).
- Jones, C. & Carvalho, L. M. V. Changes in the activity of the Madden–Julian Oscillation during 1958–2004. *J. Clim.* **19**, 6353–6370 (2006).
- Pohl, B. & Matthews, A. J. Observed changes in the lifetime and amplitude of the Madden–Julian oscillation associated with interannual ENSO sea surface temperature anomalies. *J. Clim.* **20**, 2659–2674 (2007).
- Slingo, J. M., Rowell, D. P., Sperber, K. R. & Nortley, E. On the predictability of the interannual behaviour of the Madden–Julian Oscillation and its relationship with El Niño. *Q. J. R. Meteorol. Soc.* **125**, 583–609 (1999).
- Arnold, N. P., Kuang, Z. & Tziperman, E. Enhanced MJO-like variability at high SST. *J. Clim.* **26**, 988–1001 (2013).
- Zhang, C. & Ling, J. Barrier effect of the Indo-Pacific Maritime Continent on the MJO: perspectives from tracking MJO precipitation. *J. Clim.* **30**, 3439–3459 (2017).
- Foltz, G. R. & McPhaden, M. J. The 30–70 day oscillations in the tropical Atlantic. *Geophys. Res. Lett.* **31**, L15205 (2004).
- Wheeler, M. C. & Hendon, H. H. An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. *Mon. Weath. Rev.* **132**, 1917–1932 (2004).
- Yoo, C., Feldstein, S. & Lee, S. The impact of the Madden–Julian Oscillation trend on the Arctic amplification of surface air temperature during the 1979–2008 boreal winter. *Geophys. Res. Lett.* **38**, L24804 (2011).
- Song, E. J. & Seo, K. H. Past-and present-day Madden–Julian Oscillation in CNRM-CM5. *Geophys. Res. Lett.* **43**, 4042–4048 (2016).
- Roxy, M. Sensitivity of precipitation to sea surface temperature over the tropical summer monsoon region—and its quantification. *Clim. Dyn.* **43**, 1159–1169 (2014).
- Cravatte, S., Delcroix, T., Zhang, D., McPhaden, M. & Leloup, J. Observed freshening and warming of the western Pacific warm pool. *Clim. Dyn.* **33**, 565–589 (2009).
- Dong, L. & McPhaden, M. J. The role of external forcing and internal variability in regulating global mean surface temperatures on decadal timescales. *Environ. Res. Lett.* **12**, 034011 (2017).
- Suematsu, T. & Miura, H. Zonal SST difference as a potential environmental factor supporting the longevity of the Madden–Julian Oscillation. *J. Clim.* **31**, 7549–7564 (2018).
- Sobel, A., Wang, S. & Kim, D. Moist static energy budget of the MJO during DYNAMO. *J. Atmos. Sci.* **71**, 4276–4291 (2014).
- Kim, D., Kug, J.-S. & Sobel, A. H. Propagating versus nonpropagating Madden–Julian Oscillation events. *J. Clim.* **27**, 111–125 (2014).
- Gonzalez, A. O. & Jiang, X. Winter mean lower tropospheric moisture over the Maritime Continent as a climate model diagnostic metric for the propagation of the Madden–Julian oscillation. *Geophys. Res. Lett.* **44**, 2588–2596 (2017).
- Tokina, H., Xie, S.-P., Deser, C., Kosaka, Y. & Okumura, Y. M. Slowdown of the Walker circulation driven by tropical Indo-Pacific warming. *Nature* **491**, 439–443 (2012).
- Hermes, J. C. et al. A sustained ocean observing system in the Indian Ocean for climate related scientific knowledge and societal needs. *Front. Mar. Sci.* **6**, 355 (2019).
- Subramanian, A. et al. Ocean observations to improve our understanding, modeling, and forecasting of subseasonal-to-seasonal variability. *Front. Mar. Sci.* **6**, 427 (2019).
- Vitar, F. & Robertson, A. W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Clim. Atmos. Sci.* **1**, 3 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

MJO data and identification of events

The real-time multivariate MJO (RMM) index of Wheeler and Hendon³⁰, provided by the Australian Bureau of Meteorology, is used as a preliminary reference for identifying MJO events during 1981–2018. The RMM index³⁰ relies on an Empirical Orthogonal Function analysis, which combines equatorially averaged (15° S to 15° N) lower (850 hPa) and upper (200 hPa) tropospheric zonal winds with outgoing longwave radiation (OLR, proxy indicator for convective activity). Although the RMM index efficiently captures the dominant role of zonal winds during mature phases of strong MJO events, it can be inconsistent in representing the convective conditions associated with it^{44–46}. As a result of this absence of interplay between circulation and convection, capturing the MJO events with its convective implications has been a conundrum, as the index occasionally captures non-existent events, while some events appear to occur early or late, or are even missing^{28,44–47}.

Hence, we identified MJO events by following a set of steps that consider the RMM index but clearly capture the MJO characteristics of eastward propagation and convective activity. We focus on the boreal autumn–winter–spring seasons (November–April) during which the MJO exhibits a prominent eastward propagation, and is sensitive to SST variations in the Indian and Pacific oceans⁴⁸. In order to factor in the convective activity, we used the daily OLR from the National Oceanic and Atmospheric Administration (NOAA) at 2.5° × 2.5° horizontal resolution, which has been conventionally used for detecting the MJO-related convective activity. We also verified the detected events using the high-resolution (1° × 1°) daily OLR Climate Data Record⁴⁹, which is better suited for identifying the tropical variability at sub-seasonal timescales⁵⁰. The MJO phase duration is strongly linked to the strength of MJO convection and its coupling with the largescale circulation⁵¹. Hence, the current method makes sure to capture the MJO events that exhibit strong coupling between tropical convection and largescale circulation.

The OLR on subseasonal timescales also represents other types of equatorial propagating modes of convection, such as the westward-moving equatorial Rossby waves, eastward-moving Kelvin waves and mixed Rossby–gravity waves. The MJO component is therefore filtered from the OLR data by including eastward zonal wavenumbers 1–5 and a period of 30–96 days, while the Kelvin wave component is separated by identifying eastward wavenumbers 1–14 and a period of 2–30 days, and equatorial Rossby waves by their westward zonal wavenumbers 1–10 and periods of 10–50 days^{52,53}. We select eastward propagating convective MJO events in the filtered OLR anomalies, which are initiated in the Indian Ocean (phases 1, 2 or 3)²⁸, proceed to the Pacific (phases 6, 7 or 8) and propagate through at least six of the RMM phases with an average RMM amplitude greater than one (–1.5 standard deviation). We consider the initiation date as when the RMM index indicates MJO entry into the Indian Ocean from the west and starts to propagate eastward. We find 88 such MJO events over the 38 years, during November–April. The selected events are comparable to the MJO events detected by the tracking method used by Suematsu and Miura³⁶, which is based solely on the RMM index at a threshold amplitude of 0.8, but with a relatively wide window for the band-pass filter (20–120 days).

Note that for computational purposes, the data for November–April are considered together as belonging to the initial year (for example, MJO activity during November 1981–April 1982 is considered together as representing the year 1981). Hence, although we have 38 years of data, we consider it as 37 MJO seasons.

Warm pool SST and climate data analysis

HadISST1 SST data for the period 1900–2018, obtained from the Met Office Hadley Centre, is used to estimate the changes in the Indo-Pacific warm pool and its role on the MJO phase duration. The warm pool area is estimated as the surface area covered by the climatological 28 °C

isotherm of SST, during November–April (Fig. 1), in the tropical Indo-Pacific region within 40° E to 140° W, 25° S to 25° N. To examine the state and response of atmospheric circulation to the changing SSTs and MJO, we used air temperature, winds and specific humidity values for the tropospheric column from NCEP reanalysis for the period 1981–2018 at a 2.5° × 2.5° grid resolution. The global changes in rainfall are estimated using the NOAA GPCP precipitation dataset, which combines observations and satellite precipitation data on a 2.5° × 2.5° global grid.

A breakpoint analysis⁵⁴ is conducted to identify significant shifts in the mean of the Indo-Pacific warm pool time series (Extended Data Fig. 3b). The analysis uses a Bai–Perron test⁵⁵ to determine the optimal number of breaks using Bayesian information criterion⁵⁶ and the residual sum of squares, given the minimum segment size of the time series (30 year segments used here). The location of these breakpoints can be attributed to the timing of nonlinear changes in the observed warm pool area over time. The analysis was performed using the ‘strucchange’ package in the R Statistical Software⁵⁴.

The life cycle of the MJO and the tropical ocean–atmosphere conditions are also dependent on the state of the ENSO. We use a frequency bandpass filter (2–6 years) to remove the interannual frequency band associated with ENSO-related variations, although removing all of the ENSO related variability is difficult as it can influence variability at both higher and lower frequency. The correlation analysis and trends in Figs. 3 and 4 are estimated using these filtered anomalies. The least-square linear regression and Theil–Sen slope methods are used to estimate the observed trends. The Theil–Sen approach is considered more robust than the least-squares method due to its relative insensitivity to extreme values and better performance even for normally distributed data⁵⁷.

The statistical significance of the trends, correlations and the difference of slopes⁵⁸ (Extended Data Fig. 3c) is examined using standard two-tailed Student’s *t*-tests. The significance of the trends in the time series plots are further assessed with a Mann–Kendall test with block bootstrap to validate the significance when a time series shows autocorrelation⁵⁹. Statistical significance exceeding the 95% confidence level is selected a priori as the level at which the null hypothesis can be rejected. The correlation analysis is also tested using Kendall rank correlation that is non-parametric and therefore makes no assumptions about the distribution and at the same time determine the direction and significance of the relation between the two variables⁵⁹. The correlated variables are said to be concordant if their ranks vary together (+1) and discordant if they vary differently (–1). In order to compare the differences in the distribution of the MJO phase durations in Fig. 2, we have used the Mann–Whitney *U*-test⁶⁰ to test the null hypothesis that there is no difference between two means (Extended Data Fig. 8). The Mann–Whitney *U*-test is a non-parametric test useful for relatively short time series, and also takes into account the fact that MJO variability is not normally distributed about the mean state.

Data availability

The MJO RMM index used in the study for the period 1981–2018 is available from the Australian Bureau of Meteorology (<http://www.bom.gov.au/climate/mjo/>). The monthly values of air temperature, specific humidity and winds, and the daily OLR and GPCP monthly precipitation can be obtained from the NOAA website (<https://www.esrl.noaa.gov/psd/data/gridded/>). HadISST data are available for download at the Met Office Hadley Centre website (<https://www.metoffice.gov.uk/hadobs/hadisst/>). The high-resolution daily OLR data can be acquired from the University of Maryland OLR CDR portal (<http://olr.umd.edu/>).

Code availability

The MJO events identified in this study, and the code for estimating the individual MJO phase duration and the Indo-Pacific warm pool area,

are available at <https://github.com/RoxyKoll/warmpool-mjo>. The code for filtering the MJO component from the OLR data is available from C. Schreck at GitLab (https://k3.cicsnc.org/carl/carl-ncl-tools/blob/master/filter/filter_waves.ncl).

44. Straub, K. H. MJO initiation in the real-time multivariate MJO index. *J. Clim.* **26**, 1130–1151 (2013).
45. Liu, P. et al. A revised real-time multivariate MJO index. *Mon. Weath. Rev.* **144**, 627–642 (2016).
46. Wolding, B. O. & Maloney, E. D. Objective diagnostics and the Madden–Julian oscillation. Part II: application to moist static energy and moisture budgets. *J. Clim.* **28**, 7786–7808 (2015).
47. Ventrice, M. J. et al. A modified multivariate Madden–Julian oscillation index using velocity potential. *Mon. Weath. Rev.* **141**, 4197–4210 (2013).
48. Hendon, H. H., Wheeler, M. C. & Zhang, C. Seasonal dependence of the MJO–ENSO relationship. *J. Clim.* **20**, 531–543 (2007).
49. Schreck, C., Lee, H.-T. & Knapp, K. HIRS outgoing longwave radiation—daily climate data record: application toward identifying tropical subseasonal variability. *Remote Sens.* **10**, 1325 (2018).
50. Kikuchi, K., Wang, B. & Kajikawa, Y. Bimodal representation of the tropical intraseasonal oscillation. *Clim. Dyn.* **38**, 1989–2000 (2012).
51. Seo, K.-H. & Kumar, A. The onset and life span of the Madden–Julian oscillation. *Theor. Appl. Climatol.* **94**, 13–24 (2008).
52. Wheeler, M. & Kiladis, G. N. Convectively coupled equatorial waves: analysis of clouds and temperature in the wavenumber–frequency domain. *J. Atmos. Sci.* **56**, 374–399 (1999).
53. Roundy, P. E., Schreck, C. J. III & Janiga, M. A. Contributions of convectively coupled equatorial Rossby waves and Kelvin waves to the real-time multivariate MJO indices. *Mon. Weath. Rev.* **137**, 469–478 (2009).
54. Zeileis, A., Kleiber, C., Krämer, W. & Hornik, K. Testing and dating of structural changes in practice. *Comput. Stat. Data Anal.* **44**, 109–123 (2003).
55. Bai, J. & Perron, P. Computation and analysis of multiple structural change models. *J. Appl. Econ.* **18**, 1–22 (2003).
56. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
57. Hirsch, R. M., Slack, J. R. & Smith, R. A. Techniques of trend analysis for monthly water quality data. *Wat. Resour. Res.* **18**, 107–121 (1982).
58. Cohen, P., West, S. G. & Aiken, L. S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Psychology Press, 2014).
59. Kendall, M. G. *Rank Correlation Methods* 2 edn (C. Griffin, 1948).
60. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).

Acknowledgements M.K.R. acknowledges NOAA/PMEL for the National Research Council Senior Research Associateship Award by the US National Academy of Sciences (PMEL contribution no. 4975). P.D. was supported by the ITM Research Fellowship. D.K. was supported by the DOE RGMA program (DE-SC0016223), the NOAA CVP program (NA18OAR4310300), and the KMA R&D program (KMI2018-03110). We thank N. Bond and R. Murtugudde for their comments on an early draft of this manuscript.

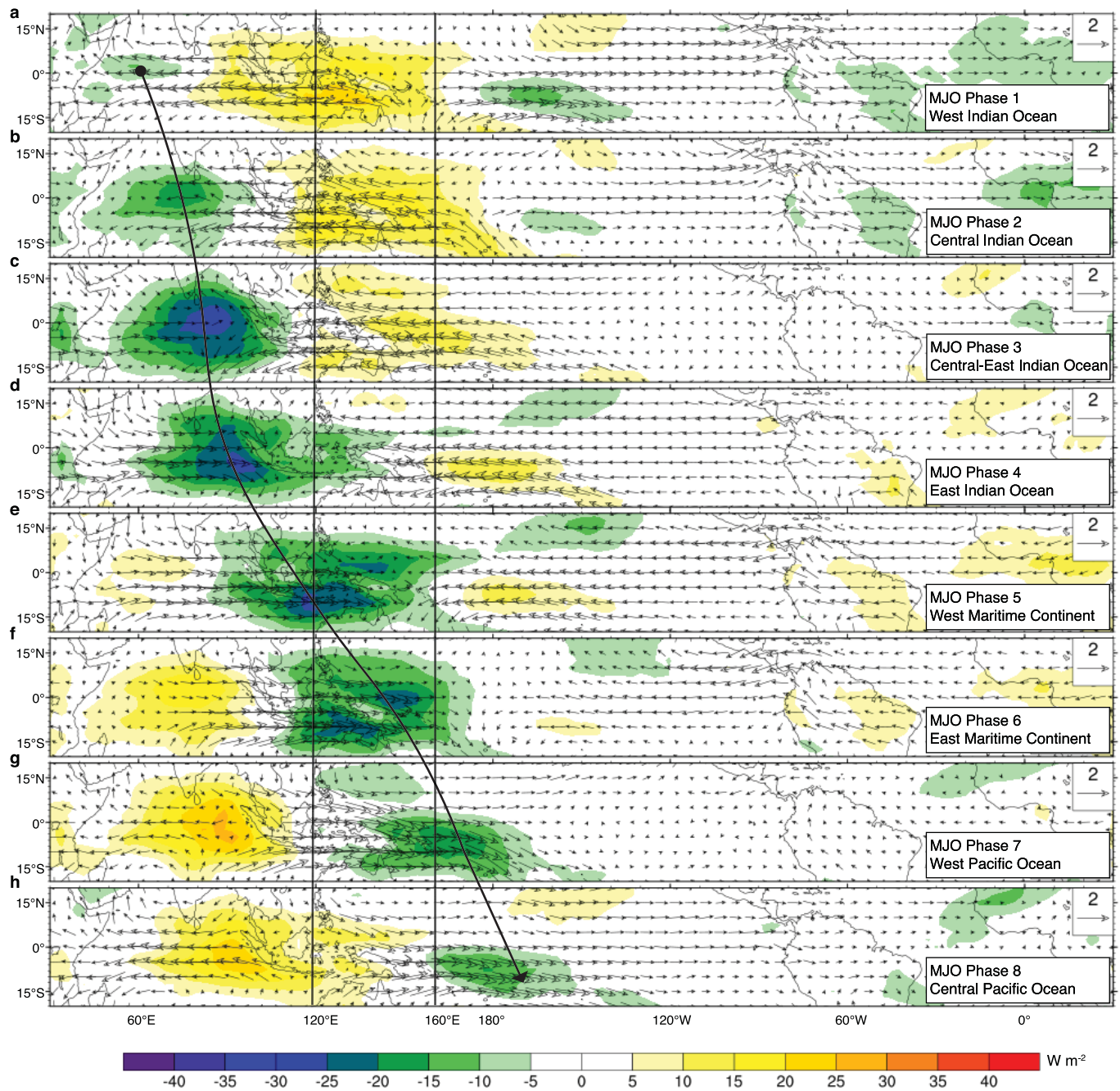
Author contributions M.K.R. conceived the study, performed the analysis and prepared the manuscript. P.D. performed the MJO detection and initial analysis. T.S. provided additional MJO tracking algorithm for verification. All co-authors contributed to the interpretation of the results and drafting of the manuscript for publication.

Competing interests The authors declare no competing interests.

Additional information

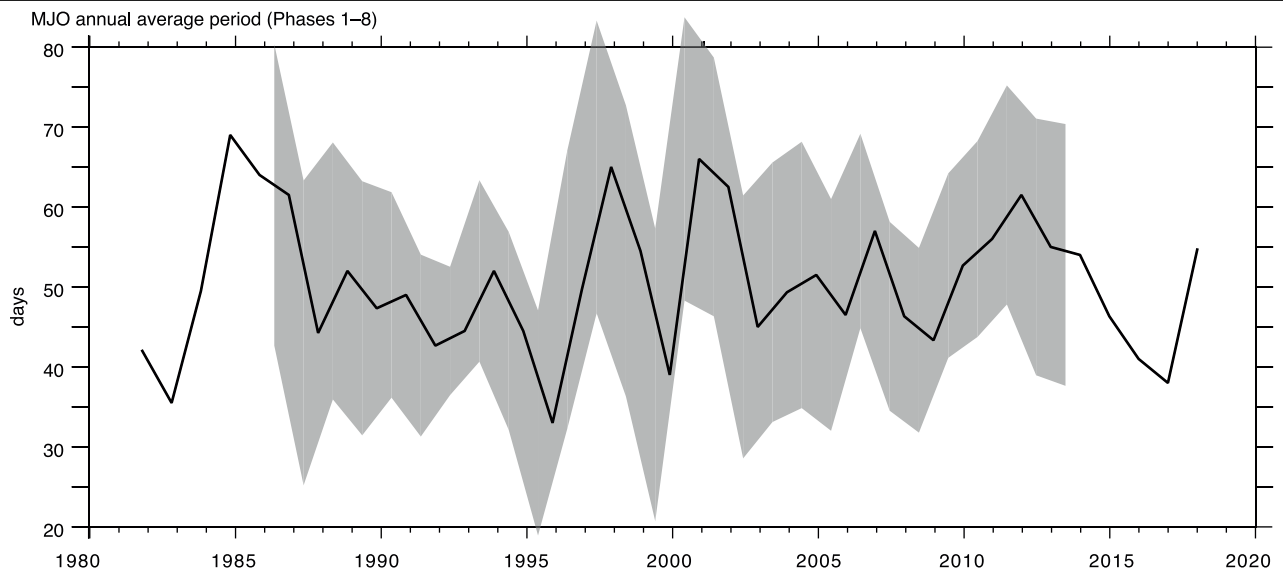
Correspondence and requests for materials should be addressed to M.K.R.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

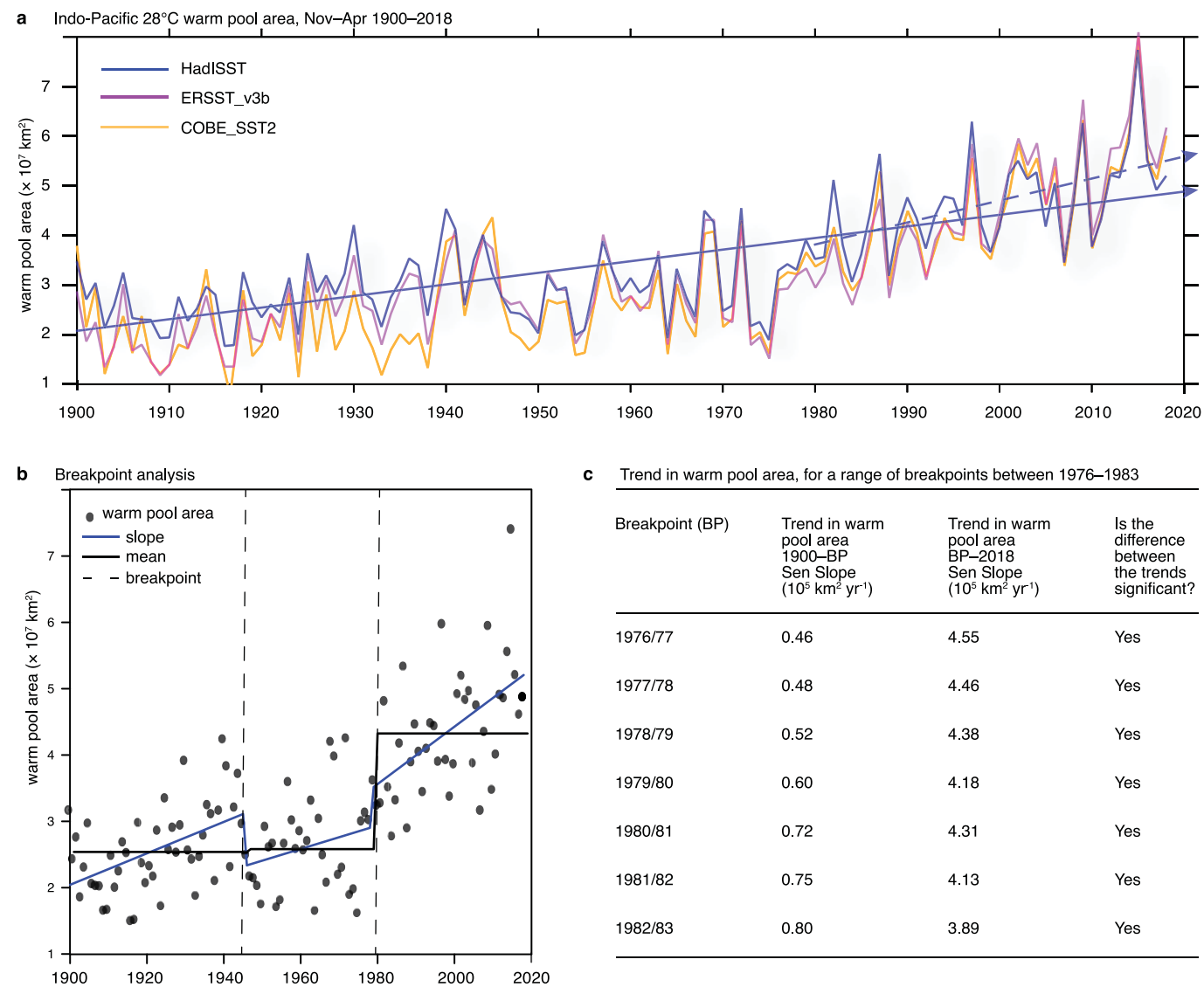


Extended Data Fig. 1 | Typical life cycle of the MJO. Composite anomalies of 30–100 day OLR (W m^{-2}) during November–April, for the period 1981–2018, showing the RMM phases 1–8. Typically, the MJO events are initiated over the Indian Ocean and move eastward over the Maritime Continent to the Pacific (a–h). The region within the solid black lines highlight the west Pacific warm

pool region (120°E to 160°E) where ocean–atmospheric changes related to the MJO lifespan are the largest. OLR values are based on the NOAA interpolated OLR dataset. The NCAR Command Language (NCL, <http://ncl.ucar.edu>) is used to plot the MJO life cycle on the map.

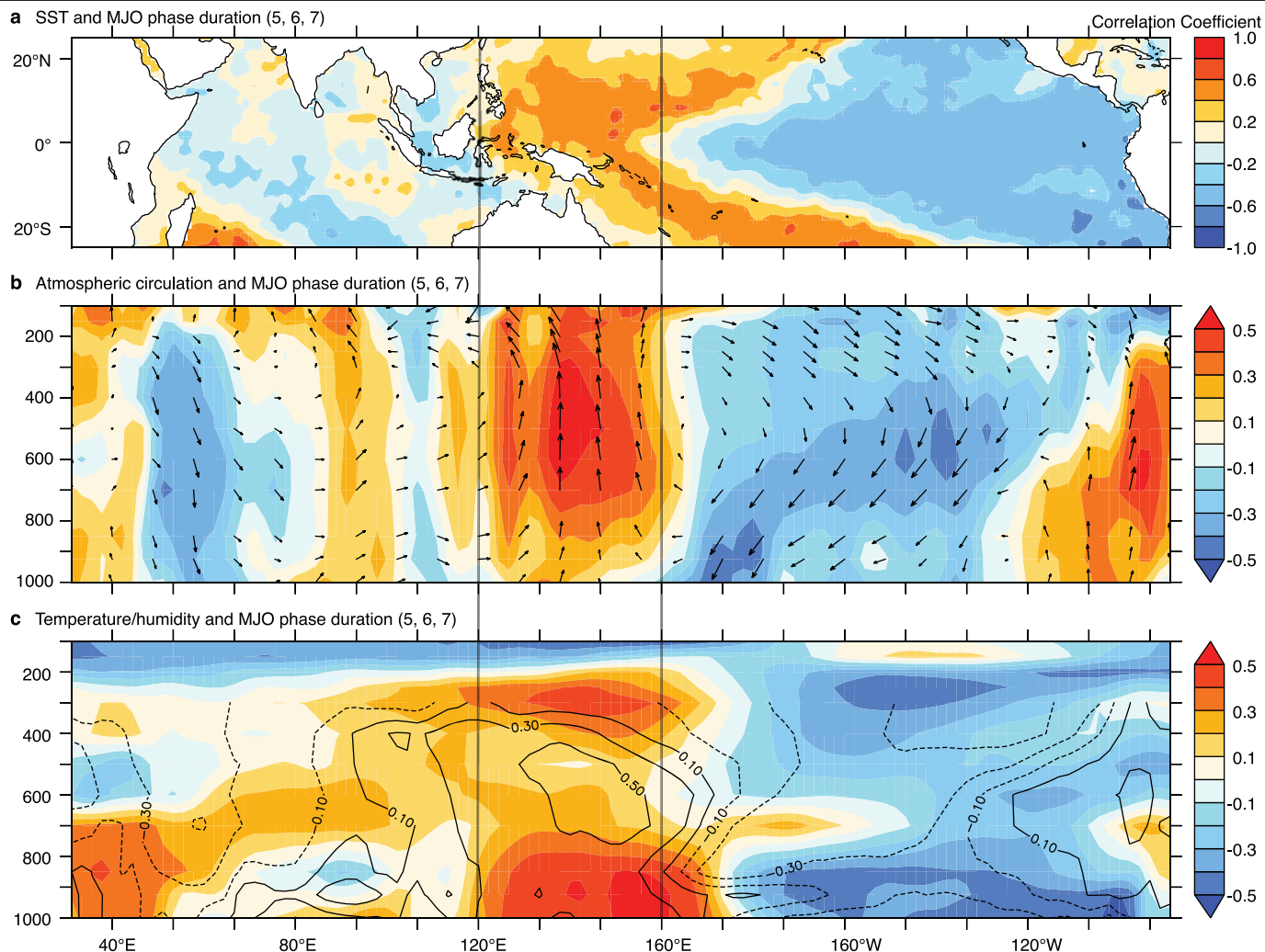


Extended Data Fig. 2 | Annual average period of MJO events. Time series of yearly average period of MJO events during November–April, 1981–2016 (phases 1–8). The grey shade overlaid on the time series represents \pm two standard deviations of the MJO phase duration over a 10-year moving window.



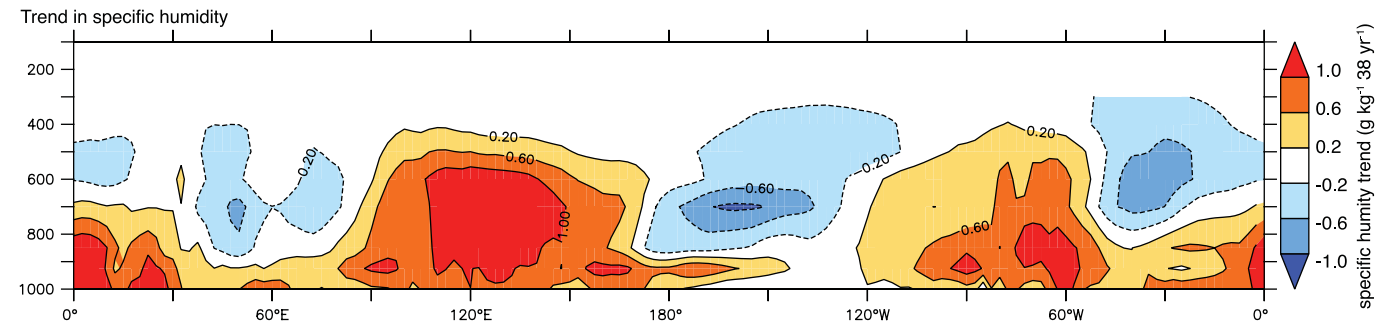
Extended Data Fig. 3 | Warm pool area in multiple datasets and breakpoint analysis. **a**, Time series of the warm pool area during November–April, 1900–2018, based on HadISST, ERSST_v3b and COBE_SST2 datasets. Theil–Sen trend estimates computed based on HadISST (as in Fig. 1) are overlaid on the time series for the entire period (solid blue line) and for 1981–2018 (dashed blue line). **b**, Breakpoint analysis identifying the significant shifts in the mean of the Indo-Pacific warm pool time series, using HadISST. The breakpoint analysis shows two shifts in the time series, the first during 1945–1946 and the second during 1979–1980. Although the rate of change in warm pool area during 1900–1945

and 1946–1979 are different, the average warm pool area remains almost the same during both the periods. The breakpoint analysis confirms that the shifts to higher warm pool values occurred in the annual series during 1979–1980. **c**, Table showing the trend in warm pool area using a range of breakpoints, from 1976–1977 to 1982–1983. The rate of warming does not change substantially with different breakpoints. At the same time, the difference between the trends is significant for all breakpoints considered. The significance of the difference between the slopes is estimated based on a *t*-test⁵⁸.



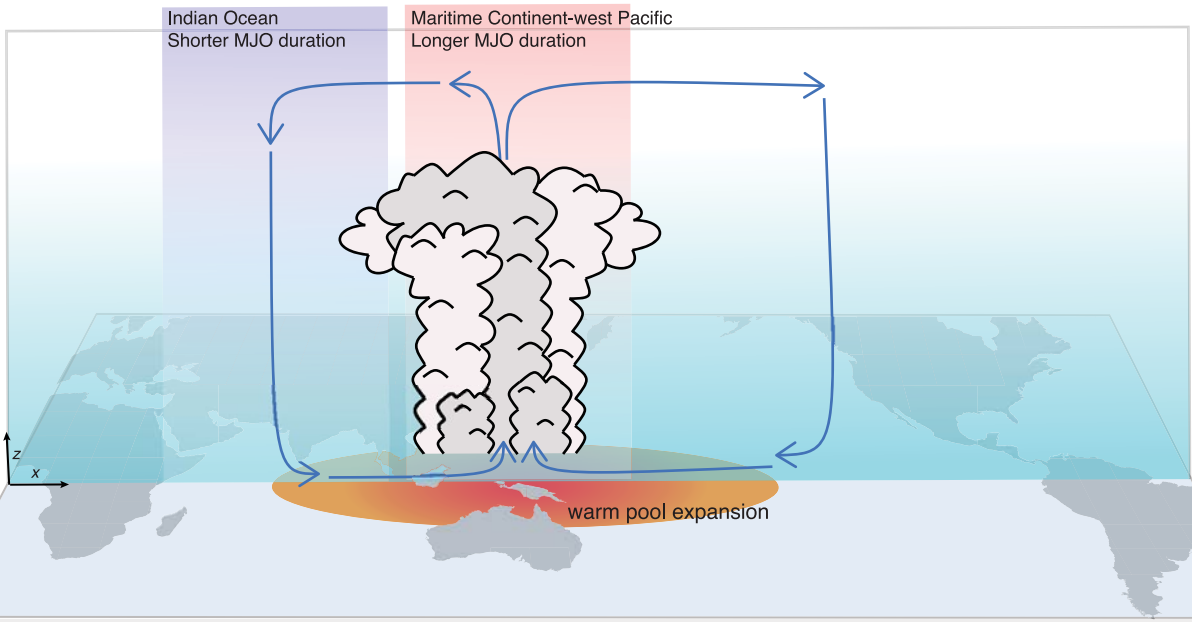
Extended Data Fig. 4 | Correlation between MJO phase duration and ocean-atmosphere conditions, without removing the trends. a–c, Correlation between yearly average of MJO phase distribution (phases 5, 6 and 7) with (a) SST anomalies, (b) winds and vertical velocity and (c) air temperature (colours) and specific humidity (contours) over the Indo-Pacific basin for November–

April, during 1981–2018 ($n = 37$). The correlation analyses are performed after removing the ENSO variability from the time series, but without removing the trends. PyFerret (<http://ferret.pmel.noaa.gov/Ferret/>) is used to generate the map and the plots.

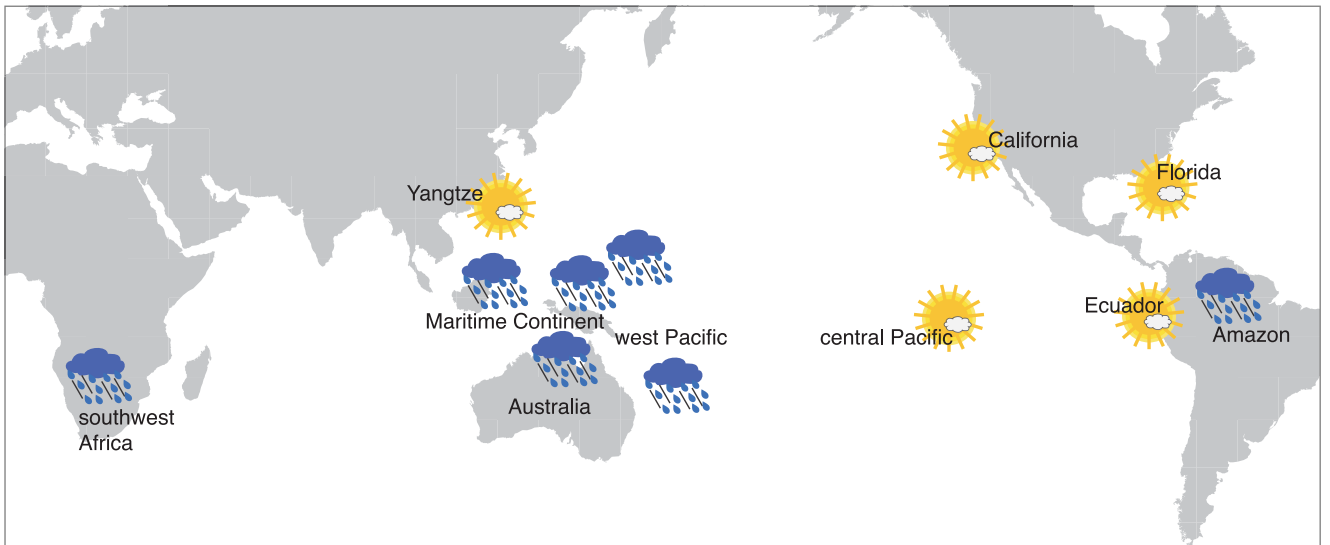


Extended Data Fig. 5 | Trend in specific humidity anomalies. Trend in specific humidity anomalies (g kg^{-1} per 38 years) for November–April, during 1981–2018. The trends indicate an increase (red colours) in tropospheric moisture over the warm pool region and a reduction (blue colours) in tropospheric moisture over the Indian Ocean (900–400 hPa levels).

a



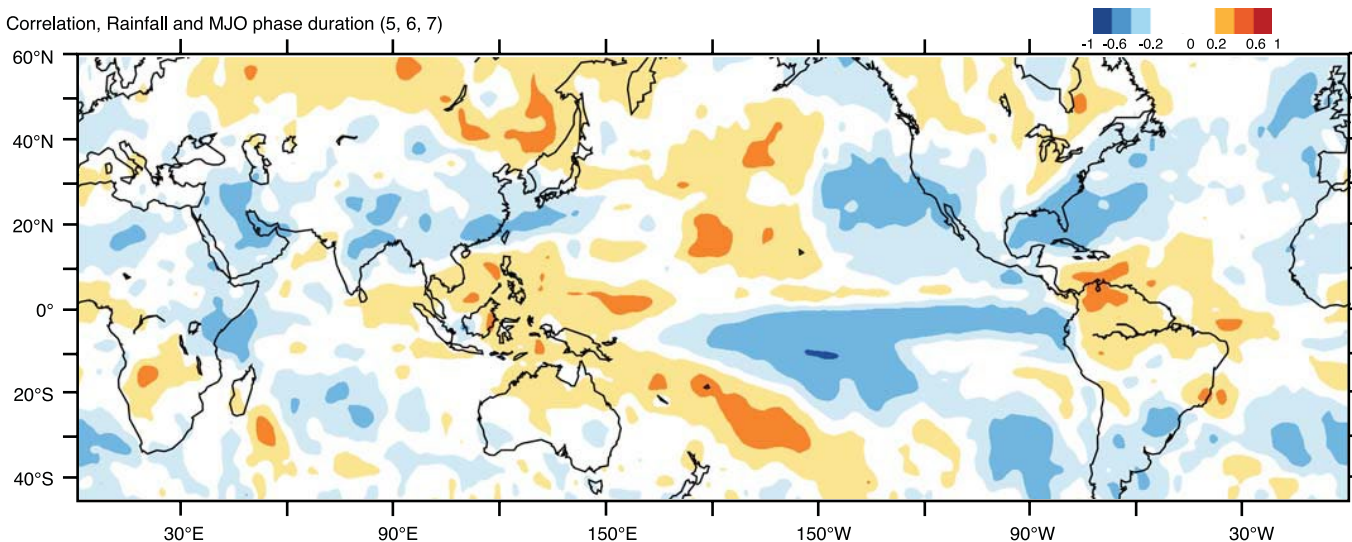
b



Extended Data Fig. 6 | Schematics showing the changes in MJO life cycle and impact on the global climate. a, As the Indo-Pacific warm pool expands with increasing SSTs, moist winds converge over the Maritime Continent-west Pacific, prolonging the MJO phase duration over this region by 5–6 days and shortening the MJO duration over the Indian Ocean by 3–4 days. **b,** As a response to the changes in the MJO phase duration, an increase in mean rainfall

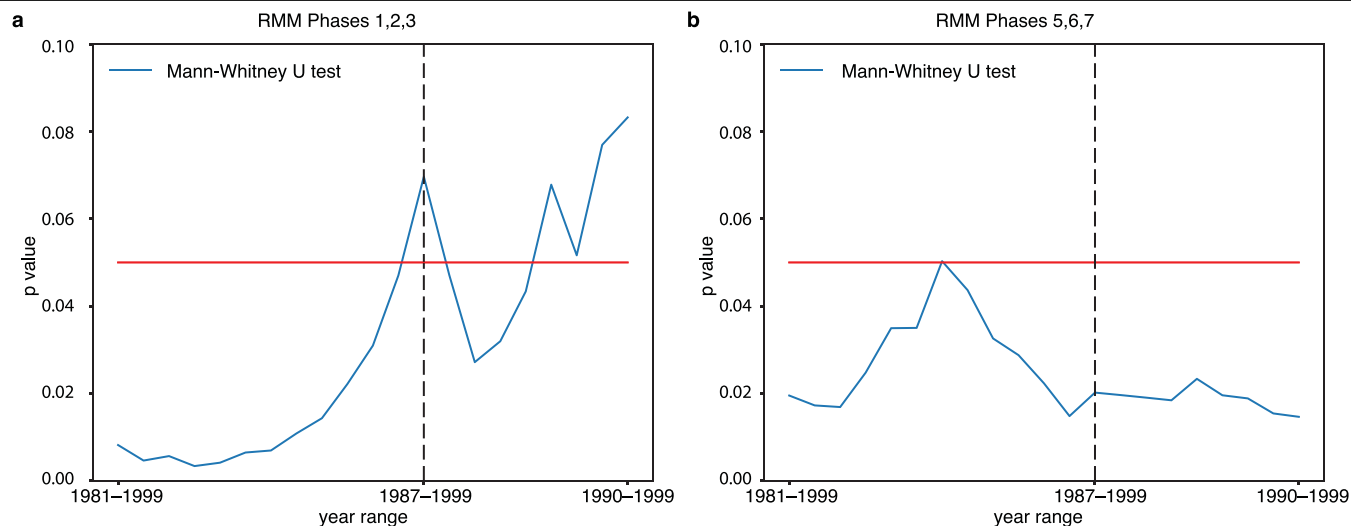
is observed over most of the Maritime Continent including southeast Asia, and over northern Australia, west Pacific, Amazon basin and southwest Africa. A decline in rainfall is observed over the central Pacific, Ecuador and California, and a slight decrease in rainfall over the Yangtze basin in China and Florida. The Generic Mapping Tools (GMT, <https://github.com/GenericMappingTools/gmt>) was used to create the map.

Correlation, Rainfall and MJO phase duration (5, 6, 7)



Extended Data Fig. 7 | Relationship between MJO phase duration and global rainfall, without removing the trends. Correlation between the MJO phase duration (phases 5, 6 and 7) and rainfall anomalies for November–April, during 1981–2018. The correlation analysis is performed after removing the ENSO

variability from the time series, but without removing the trends. Rainfall values are based on the GPCP dataset. The Generic Mapping Tools (GMT, <https://github.com/GenericMappingTools/gmt>) was used to create the map.



Extended Data Fig. 8 | Mann-Whitney *U*-test for testing the significance of the differences in MJO phase duration. The difference in the mean of the MJO phase duration distributions is tested for different starting points. The *P* values are computed for different groups (1981–1999, 1982–1999 to 1990–1999) as the first sample and 2000–2018 as the second sample. **a**, According to the Mann–Whitney *U*-test, the difference in MJO phase duration (1, 2, 3) is statistically

robust ($P < 0.05$, where we can reject the null hypothesis) for the most part of the varying first sample (1981–1999 to 1990–1999, except 1987–1999 where $P = 0.07$). **b**, For the MJO phase duration (5, 6, 7) the difference in mean is always statistically robust (where we can reject the null hypothesis) for the varying first sample (1981–1999 to 1990–1999, where P always < 0.05).

Genetic predisposition to mosaic Y chromosome loss in blood

<https://doi.org/10.1038/s41586-019-1765-3>

Received: 9 January 2019

Accepted: 4 October 2019

Published online: 20 November 2019

Deborah J. Thompson¹, Giulio Genovese^{2,3,4}, Jonatan Halvardson⁵, Jacob C. Ulirsch^{3,6}, Daniel J. Wright^{7,8}, Chikashi Terao^{9,10,11,12}, Olafur B. Davidsson¹³, Felix R. Day^{7,14}, Patrick Sulem¹³, Yunxuan Jiang¹⁵, Marcus Danielsson⁵, Hanna Davies⁵, Joe Dennis¹, Malcolm G. Dunlop¹⁶, Douglas F. Easton¹, Victoria A. Fisher¹⁷, Florian Zink¹³, Richard S. Houlston¹⁸, Martin Ingelsson¹⁹, Siddhartha Kar²⁰, Nicola D. Kerrison⁷, Ben Kinnarsley¹⁸, Ragnar P. Kristjansson¹³, Philip J. Law¹⁸, Rong Li²¹, Chey Loveday¹⁸, Jonas Mattisson⁵, Steven A. McCarroll^{2,3,4}, Yoshinori Murakami²², Anna Murray²³, Pawel Olszewski²⁴, Edyta Rychlicka-Buniowska^{5,24}, Robert A. Scott⁷, Unnur Thorsteinsdottir^{13,25}, Ian Tomlinson²⁶, Behrooz Torabi Moghadam⁵, Clare Turnbull^{18,27}, Nicholas J. Wareham⁷, Daniel F. Gudbjartsson^{13,28}, International Lung Cancer Consortium (INTEGRAL-ILCCO)²⁹, The Breast Cancer Association Consortium²⁹, Consortium of Investigators of Modifiers of BRCA1/2²⁹, The Endometrial Cancer Association Consortium²⁹, The Ovarian Cancer Association Consortium²⁹, The Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) Consortium²⁹, The Kidney Cancer GWAS Meta-Analysis Project²⁹, eQTLGen Consortium²⁹, Biobank-based Integrative Omics Study (BIOS) Consortium²⁹, 23andMe Research Team²⁹, Yoichiro Kamatani^{9,12,30}, Eva R. Hoffmann³¹, Steve P. Jackson^{32,33}, Kari Stefansson^{13,25}, Adam Auton¹⁵, Ken K. Ong⁷, Mitchell J. Machiela¹⁷, Po-Ru Loh^{3,34}, Jan P. Dumanski^{5,24}, Stephen J. Chanock¹⁷, Lars A. Forsberg^{5,35,36} & John R. B. Perry^{7,14,36*}

Mosaic loss of chromosome Y (LOY) in circulating white blood cells is the most common form of clonal mosaicism^{1–5}, yet our knowledge of the causes and consequences of this is limited. Here, using a computational approach, we estimate that 20% of the male population represented in the UK Biobank study ($n = 205,011$) has detectable LOY. We identify 156 autosomal genetic determinants of LOY, which we replicate in 757,114 men of European and Japanese ancestry. These loci highlight genes that are involved in cell-cycle regulation and cancer susceptibility, as well as somatic drivers of tumour growth and targets of cancer therapy. We demonstrate that genetic susceptibility to LOY is associated with non-haematological effects on health in both men and women, which supports the hypothesis that clonal haematopoiesis is a biomarker of genomic instability in other tissues. Single-cell RNA sequencing identifies dysregulated expression of autosomal genes in leukocytes with LOY and provides insights into why clonal expansion of these cells may occur. Collectively, these data highlight the value of studying clonal mosaicism to uncover fundamental mechanisms that underlie cancer and other ageing-related diseases.

Each day the human body produces billions of highly specialized blood cells, which are generated from a self-renewing pool of 50,000–200,000 haematopoietic stem cells (HSCs)⁶. As these cells age and divide, mutations and errors in mitosis create genetic diversity within the cells of the HSC pool and their progenitors. If a genetic alteration confers a selective growth advantage to one cell over the others, clonal expansion may occur. This process propels the lineage to a disproportionately high frequency, creating a genetically distinct sub-population of cells. In the literature this is commonly referred to as clonal haematopoiesis, or, more broadly (that is, not restricted to the context of leukocytes), clonal mosaicism⁷ or aberrant clonal expansion⁵.

Population-based studies assessing the magnitude and effect of clonal mosaicism have been largely limited by the challenges of

accurately detecting the expected low-cell-fraction mosaic events in leukocytes using genotype-array or sequence read data⁸. Advances in statistical methodology have improved sensitivity, and some approaches are now able to catalogue mosaic events at higher resolution across the genome^{9,10}. Large structural mosaic events that are detected can vary considerably in size—from 50 kb to entire chromosomes in length—and are typically present in only a small fraction of circulating leukocytes (less than 5%). It is well established that loss of the sex chromosomes, particularly the Y chromosome (LOY) in men, is by far the most frequently observed somatic change in leukocytes^{1,2,11}. However, it remains unclear whether and why the absence of a Y chromosome provides a selective growth advantage in these cells; we hypothesize that this could be due to the loss of a putative Y-linked

A list of affiliations appears at the end of the paper.

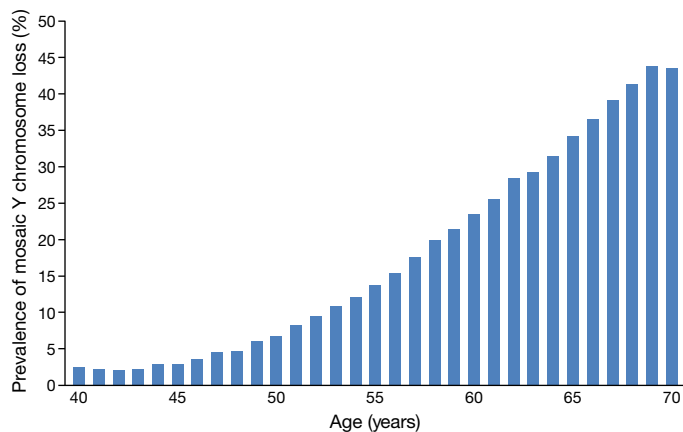


Fig. 1 | Prevalence of mosaic Y chromosome loss by age in male participants in the UK Biobank study. Bar chart shows the full age distribution of all male participants in the UK Biobank study ($n = 205,011$) at baseline.

cell-growth suppressor gene; loss of a Y-linked transcription factor that influences the expression of cell-growth-related autosomal genes; or the reduced energy cost of cellular divisions.

Our understanding of why some individuals, but not others, exhibit clonal mosaicism in the blood is also limited. Previous studies have demonstrated robust associations between clonal mosaicism and age, sex (clonal mosaicism is more frequent in males), smoking and inherited germline genetic predisposition^{3,4,7,8,12–15}. Epidemiological studies have challenged the view that LOY in the haematopoietic system is a phenotypically neutral event: epidemiological associations have been observed with various forms of cancer^{3,16–20}, autoimmune conditions^{21,22}, age-related macular degeneration²³, cardiovascular disease²⁴, Alzheimer's disease²⁵, type 2 diabetes¹⁵, obesity¹⁵ and all-cause mortality^{15,16}. The extent to which such observations represent causal association, reverse causality or confounding factors is unclear. Furthermore, if these do represent causal effects, the mechanisms that underlie such effects are unknown.

Key questions are whether the loss of a Y chromosome from circulating leukocytes has a direct functional effect (for example, impairs immune function) and whether LOY in leukocytes is a barometer of broader genomic instability in leukocytes and other cell types. Understanding the mechanisms that drive clonal mosaicism and identifying genes that impart a proliferative advantage to cells may help to answer these questions and provide insights into the mechanisms that underlie diseases of ageing. To this end we sought to identify novel susceptibility loci for LOY—a form of clonal mosaicism that is relatively easy to detect and has high prevalence in the male population. Previous genome-wide association studies (GWASs) for LOY identified 19 common susceptibility loci and highlighted the relevance of LOY as a biomarker of cell-cycle efficiency and the DNA damage response in leukocytes^{3,4}. Here, we adapt a previously described computational approach¹⁰ to detect LOY in over 200,000 men from the UK Biobank study. We identify 137 novel loci that we use, along with the known 19 loci⁴, to demonstrate a shared genetic architecture between LOY, susceptibility to non-haematological types of cancer and reproductive ageing in women. These data, in aggregate, support the hypothesis that LOY in leukocytes is a biomarker of genomic instability in other cell types, with functional consequences across diverse biological systems.

Estimating mosaic Y chromosome loss

Previous studies assessing LOY used a quantitative measure derived from the mean \log_2 -transformed R ratio of signal intensity (mLRR-Y)—that is, the ratio of observed to expected normalized-intensity values (R) of all array-genotyped single-nucleotide polymorphisms (SNPs) in

the non-pseudo-autosomal regions (non-PARs) of the Y chromosome. Here, we adapted a long-range phasing approach for the detection of mosaic events¹⁰ to estimate a dichotomous classification, which uses allele-specific genotyping intensities in the pseudo-autosomal region (PAR); we term this measure PAR-LOY (see Methods). This was applied to 205,011 men from the UK Biobank study (aged 40–70), among whom we identified 41,791 (20%) with detectable LOY. Men with LOY had an mLRR-Y score (derived using variants outside of the PAR) 0.9 standard deviations lower on average (95% confidence interval 0.88–0.91) than men without LOY (mean mLRR-Y score of -0.046 versus 0.009), reflecting the expected lower level of intensity that would result from a reduction in Y chromosomal genetic material. Consistent with previous observations of clonal mosaicism, current smokers were at a higher risk of LOY (odds ratio (OR) 1.62 (95% confidence interval 1.57–1.66)) and there was a strong association with age, with the prevalence increasing from 2.5% at age 40 to 43.6% at age 70 (Fig. 1).

Identifying genetic determinants

We estimated a heritability of 31.7% (95% confidence interval 29.9–33.6%) for LOY, distributed across all individual chromosomes in proportion to their relative sizes (Extended Data Fig. 1). To identify individual genetic variants that underlie this heritability, we performed a GWAS for LOY and identified 18,146 variants with genome-wide significant associations ($P < 5 \times 10^{-8}$). We resolved these into 156 statistically independent signals (Supplementary Table 1), which included all 19 loci that were previously reported⁴. Effect sizes for these 156 associations ranged from an odds ratio of 1.03 to 2.02, with frequencies of LOY risk alleles between 0.25% and 99.8% (Extended Data Fig. 2). An analysis in which men with any past or current diagnosis of cancer were excluded showed no change in effect estimates across the 156 loci (Extended Data Fig. 3), and a drop in the mean χ^2 value that was proportionate to the reduction in sample size.

We directly compared the power of our PAR-LOY calls to the previously used mLRR-Y-derived measures by performing an mLRR-Y-based GWAS in the same study samples (Supplementary Table 1). Across the 156 loci we observed an average increase of around 2.5-fold in the χ^2 association statistic using the PAR-LOY approach, which is exemplified by the strongest-associated variant (rs17758695 (*BCL2*)) increasing in significance from $P_{\text{mLRR-Y}} = 7.5 \times 10^{-65}$ to $P_{\text{PAR-LOY}} = 4.1 \times 10^{-147}$. Only 61 of the 156 loci would have reached genome-wide significance in an analysis that was based on mLRR-Y. Across the genome, the genomic inflation factor λ_{GC} (that is, the ratio of the expected to observed median test statistic) increased from 1.15 to 1.20 (mean χ^2 value from 1.28 to 1.47), with no evidence of signal inflation due to population structure (linkage-disequilibrium score regression intercept 1.01). Simulation analyses demonstrated that the power of PAR-LOY over mLRR-Y depends on both the sample size and the ratio of genotyped PAR1 to non-PAR SNPs on the Y chromosome (Methods, Extended Data Fig. 4).

To confirm the validity of our identified signals we sought replication in three independent datasets. First, we used data from 653,019 male research participants from the personal genetics company 23andMe, Inc. (Supplementary Table 1). These samples differed from the discovery samples both in terms of the source of the DNA (saliva rather than peripheral blood) and the type of LOY measurement (quantitative mLRR-Y rather than dichotomous PAR-LOY calls). Despite this heterogeneity, all but one of the 154 loci (2 failed quality control) had directionally concordant effects ($P = 1.4 \times 10^{-44}$, binomial sign test), with 126 exhibiting a nominally significant association ($P < 0.05$) and 88 at a more-conservative threshold ($P < 0.05/156$). Second, we sought further confirmation from the Icelandic deCODE study ($n = 8,715$) in which LOY was estimated using sequence reads from whole-genome sequencing (DNA extracted from blood), rather than array data. These data demonstrated an overall directional consistency of 94% across the associated loci (140 out of 149 variants tested; $P = 2.3 \times 10^{-31}$, binomial

sign test) and 74 nominally significant associations (Supplementary Table 1). Third, we replicated our loci in a set of 95,380 men of Japanese ancestry from the Biobank Japan project, in which LOY was estimated using mLRR-Y in whole blood. Of the 100 variants out of 156 that passed quality control and were polymorphic in East Asians, 92 had a consistent direction of effect ($P = 3.2 \times 10^{-19}$, binomial sign test). Of these, 29 reached genome-wide significance in these data alone and 73 had at least nominally significant association (Supplementary Table 1).

Finally, as a negative control we performed an analysis of mLRR-Y scores estimated in 245,349 women from the UK Biobank study (Supplementary Table 1), reflecting experimental noise in intensity variation. This did not produce any significant associations after Bonferroni correction across the 156 loci ($P_{\max} = 4.3 \times 10^{-3}$). In aggregate, these data strongly suggest that our discovery analysis identifies genetic determinants of LOY that are robust to ancestry, measurement technique and DNA source.

Implicated genes, cell types and pathways

We used various approaches to move from genomic association to identifying potentially causal variants, functional genes, cell types and biological pathways that are associated with LOY (Methods). First, we performed Bayesian fine-mapping (Methods) to quantify the probability that any single variant at a locus was causal for LOY by disentangling the effects of linkage disequilibrium (Extended Data Fig. 5, Supplementary Tables 2, 3). Fine-mapping identified at least one variant with reasonable confidence (posterior probability greater than 10%) in 80% (101 out of 126) of regions, including at least one very-high-confidence variant (posterior probability greater than 75%) in 25% (31 out of 126) of regions (Extended Data Fig. 5). These variants were enriched in exons of protein-coding genes, their promoters, their transcribed but untranslated regions, and in haematopoietic regulatory regions marked by accessible chromatin (Extended Data Fig. 5, Supplementary Table 4).

Using both fine-mapped variants and genome-wide polygenic signal (Methods), we found that haematopoietic stem and progenitor cells (HSPCs) were the most strongly enriched cell types for variants associated with LOY (Extended Data Figs. 5, 6, Supplementary Tables 5, 6). Among the fine-mapped variants, we further subdivided this enrichment into three distinct temporal modes that are indicative of an increasing regulatory capacity across haematopoiesis (Extended Data Fig. 5). These observations suggest that many of our identified variants exert their effects directly in haematopoietic stem cells, rather than more-differentiated types of white blood cells. This is in stark contrast to variants associated with the production of terminal types of blood cells, which are enriched at terminal blood progenitors and depleted in HSPCs²⁶.

We next used two approaches (Methods) to map associated genetic variants to genes through expression effects (using expression quantitative trait loci; eQTLs) in whole blood, which implicated a total of 110 unique transcripts (Supplementary Tables 8–10). This included the *HLA-A* gene—our lead variant in this region (6:29835518_T_A) tagged the HLA-A*02:01 allele (Supplementary Table 11). We also identified genes that contain a non-synonymous variant that either fine-mapped (posterior probability greater than 10%) or was in high linkage disequilibrium ($r^2 > 0.8$) with an index variant, highlighting 22 genes (Supplementary Table 8).

An analysis of biological pathways using two approaches (Methods) identified a number of associated pathways, the majority of which converged on aspects of cell-cycle regulation and the DNA damage response (Supplementary Tables 12, 13).

Overlap with cancer susceptibility loci

Although detectable clonal mosaicism is clearly associated with future risk of haematological cancers¹⁰, its relationship with other cancers is

less clear. Using data curated by the Open Targets platform and gene set enrichment analysis (Methods), we found that LOY-associated variants were preferentially found near genes involved in cancer susceptibility ($P = 9.9 \times 10^{-7}$), somatic drivers of tumour growth ($P = 7 \times 10^{-4}$) and targets of cancer therapies that are approved or in trial ($P = 0.05$). In total, 18 out of the 156 mosaic LOY-associated variants were correlated ($r^2 > 0.1$) with known susceptibility variants for one or more type of non-haematological cancer (Supplementary Table 14), including cancer of the breast, prostate, testicles, kidney, skin (melanoma) and brain. Notable examples include a loss-of-function variant in the checkpoint kinase 2 gene (*CHEK2*) (rs186430430 $r^2 = 1$ with frameshift variant 1100delC), which confers an approximately 2.3-fold risk of breast cancer²⁷, and an intronic signal (rs56345976) in the telomerase reverse transcriptase gene (*TERT*), which is in modest linkage disequilibrium ($r^2 = 0.12$) with variants associated with longer telomeres and with increased risks of breast, ovarian and prostate cancers and glioblastoma, but is also seen to be protective in other cancers²⁸.

To systematically assess the relationship between LOY susceptibility and cancer risk, we tested a genetic risk score (Methods) composed of our 156 variants on two male-specific cancers (Fig. 2, Supplementary Table 15). Genetically predicted LOY was associated with increased risk of both prostate cancer (OR 1.68 (95% confidence interval 1.33–2.11); $P = 1.9 \times 10^{-5}$) and testicular germ cell tumour (OR 2.97 (1.45–6.07); $P = 0.003$). Additional publicly available GWAS data for cancers in both sexes showed directionally consistent associations for glioma (OR 2.36 (1.34–4.17); $P = 0.004$), renal cell carcinoma (OR 2.00 (1.24–3.21); $P = 0.005$), lung cancer (OR 1.28 (0.98–1.68); $P = 0.07$), colorectal cancer (OR 1.18 (0.93–1.50); $P = 0.16$) and chronic lymphocytic leukaemia (OR 1.27 (0.75–2.16); $P = 0.37$) (Fig. 2, Supplementary Table 15).

Relevance to health outcomes in women

Mosaic LOY in blood cells has been epidemiologically associated with a broad range of diseases. If this link is causal, it is probably explained by one (or both) of two mechanisms: either LOY in leukocytes has a direct physiological effect—for example, through impaired immune function—and/or it acts as a barometer by providing a readily detectable manifestation of genomic instability that occurs in parallel in other tissues. In an ideal scenario, this question would be addressed by assessing clonal mosaicism in studies of large populations, with DNA being extracted from a broad range of cell and tissue types. However, in the absence of such a study, we hypothesized that testing the relevance of our identified LOY-associated variants in women would be informative—any association between the two could not be explained by a direct effect of LOY, given that females have no Y chromosome.

To assess this, we tested a polygenic risk score made up of our 156 lead variants for association with three female-specific cancers: breast, endometrial and ovarian cancer (Fig. 2, Supplementary Table 15). We observed a significant association with breast cancer (OR 1.25 (1.04–1.49); $P = 0.016$) and directionally consistent results in the smaller studies of endometrial (OR 1.18 (0.94–1.48); $P = 0.14$) and ovarian (OR 1.02 (0.81–1.30), $P = 0.86$) cancers.

We next tested the same score on a female-specific non-cancer trait that is also underpinned by genomic instability—age at natural menopause. Previous studies in humans and animals have shown that age at menopause is substantially biologically determined by the ability of oocytes to detect, repair and respond to DNA damage^{29,30}. We found that genetically increased risk of LOY was associated with later age at menopause ($P = 0.003$, Supplementary Table 16), with the *CHEK2* locus individually reaching genome-wide significance for association with menopause ($P = 7.9 \times 10^{-22}$). A repeated analysis of a genetic risk score that excluded *CHEK2* retained significance ($P = 0.017$).

Given the observation that genetic susceptibility to LOY in leukocytes is affecting broader biological systems in these women, it is reasonable to speculate that actual LOY in leukocytes in men similarly represents

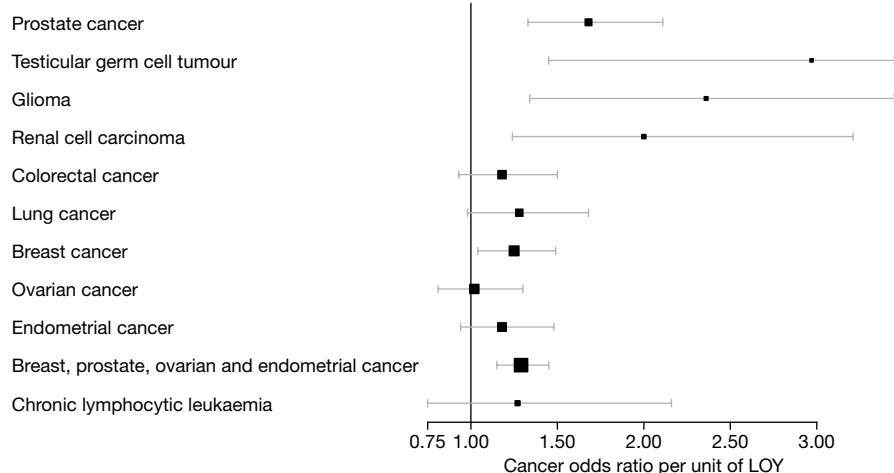


Fig. 2 | Influence of genetic susceptibility to LOY on cancer outcomes. The genetic risk score comprises the 156 LOY-associated loci identified in the UK Biobank discovery analysis ($n = 205,011$). Squares are proportional to the

sample size. Error bars denote 95% confidence intervals around the point estimate OR effect, with the value 1 denoting no effect.

a biomarker of genomic instability occurring in other cell and tissue types.

The effect of LOY in a single cell

To help understand whether—and if so, why—LOY may provide a growth advantage to a cell, and the potential mechanisms that link LOY to disease, we used the 10X Genomics Chromium Single Cell 3' platform for single-cell RNA sequencing (scRNA-seq). scRNA-seq was performed on peripheral blood mononuclear cells (PBMCs) that were collected from 19 male donors (aged 64–89) who were not selected on the basis of any measure of clonal mosaicism. After standard quality control steps (Methods), we sequenced and profiled gene expression across 86,160 single cells. Under normal conditions, blood cells express a set of genes located in the male-specific region of the Y chromosome. The LOY status of individual cells could therefore be determined by the absence of expression from these genes, which we identified in 13,418 of the cells (15.6% across all cells, ranging from 7 to 61% within individuals).

We next tested whether any of the genes that were identified in our LOY GWAS were differentially expressed between cells with and without the Y chromosome (Extended Data Fig. 7). This analysis highlighted *TCL1A* (mapped LOY locus rs2887399, 162 bp away), with the allele that conferred an increased risk of LOY being associated with higher expression of *TCL1A* in blood cells (Supplementary Table 10). The single-cell data showed that among the major types of leukocytes, *TCL1A* was expressed only in B lymphocytes (Extended Data Fig. 7), and LOY was detected in 11.3% of these cells (ranging from 2% to 56% within individuals). B lymphocytes without a Y chromosome ($n = 277$ cells) had 75% higher normalized expression of *TCL1A* than those with a Y chromosome ($n = 2,459$; fold change 1.75; $P < 0.0001$ (Wilcoxon test in Seurat)). We also performed an in-house resampling test to evaluate this difference and validated a substantial upregulation of *TCL1A* in cells with LOY (fold change 1.68; $P < 0.0001$) (Extended Data Fig. 7). An analysis within each individual showed that single cells with LOY had consistently higher expression of *TCL1A*—ruling out any bias by *TCL1A* genotype (Extended Data Fig. 8).

To evaluate the magnitude of the 75% overexpression of the *TCL1A* gene in B lymphocytes with LOY, we compared the expression changes of other genes proximal to our identified GWAS loci. Of the genes that we prioritized at each of our GWAS loci ('consensus genes'; Supplementary Table 8), 71 were expressed in over 5% of the B lymphocytes and included in the comparison, but only *TCL1A* demonstrated a significant fold change in expression (Extended Data Fig. 7).

TCL1A encodes T cell leukaemia/lymphoma protein 1A, which functions as a co-activator of the cell survival kinase AKT, and is often over-expressed in haematological malignancies of T and B cells³¹. These data provide a possible explanation for the growth advantage conferred to cells that are missing a Y chromosome. The independent effect of *TCL1A* genotype also suggests a possible bidirectional involvement for *TCL1A*. Ultimately, further experimental work will be required to fully elucidate the aetiological implications of altered *TCL1A* expression in these cells.

Discussion

This study provides several advances in our understanding of the biology that may underlie mosaic Y chromosome loss in circulating leukocytes, and its probable consequences. Our improved ability to detect LOY and increased sample size led to an eightfold increase in the number of associated genetic determinants, and enabled us to make several observations.

The origin of LOY at the level of a single cell is perhaps most readily explained by chromosome mis-segregation events during mitosis. Consistent with this, many of the identified loci are proximal to genes involved in key mitotic processes (Extended Data Fig. 9)—notably genes encoding central components of the condensin complex (for example, *NCAPG2* and *SMC2*), which affect the structure of mitotic chromosomes³²; *CENPN*, *CENPU*, *PMF1* and *ZWILCH*, which are involved in the assembly, structure and function of the kinetochore; and *SPDL1*, which is involved in spindle formation, that together form the main machinery of chromosome congression and segregation^{33,34}. The proteins encoded by *MAD2L1* (alongside *MAD1L1* and *MAD2L1BP*) and *ZWILCH* are core components of the mitotic spindle-assembly checkpoint³⁵, which ensures that chromatids are bi-oriented at the metaphase plate and under bipolar tension before disinhibiting the anaphase-promoting complex (of which *ANAPC5* is a component) to allow progression from metaphase. Many genes governing wider cell-cycle progression, including cyclins (*CCND2* and *CCND3*), regulators of cyclin (*CDKN1B*, *CDKN1C* and *CDK5RAP1*) and major checkpoint kinases (*ATM*) are also identified here, emphasizing the importance of processes across the cell cycle in determining LOY. Some of the remaining genes that we identify encode proteins that are involved in sensing and responding to DNA damage (*SETD2*, *DDI2*, *PARP1*, *ATM*, *TP53* and *CHEK2*) and apoptotic processes (*PMAIP1*, *SPOP*, *LTBR*, *SGMS1*, *TP53INP1* and *DAP*). The BCL-2 family, a conserved set of proteins that regulate caspase-mediated apoptosis by controlling mitochondrial release of cytochrome *c*, are also particularly well represented (*BCL2*, *BAX*, *BCL2L1* and *BCL2L11*)³⁶. These

themes are consistent with the hypothesis that secondary to the initial mis-segregation event, clonal expansion of cells with LOY requires an environment permissive to proliferation of aneuploid cells, in which the normal processes that detect and terminate these cells are avoided.

A link between LOY and cancer susceptibility seems plausible conceptually, given the nature of the genes identified. We found a substantial overlap of LOY-associated variants with known cancer susceptibility loci, somatic drivers of tumour growth and genes targeted by licensed or in-trial cancer therapeutics. A notable example is *PARP1* (the target of PARP inhibitors). In this case, the lead SNP is highly correlated with a missense variant (V762A), the minor allele for which (the alanine substitution) is protective against LOY and has experimentally been shown³⁷ to reduce the catalytic activity of PARP1 by 30–40%. More broadly, we found evidence for a systematic relationship between genetic susceptibility to LOY and risk of breast, prostate, testicular and renal cell carcinomas (Fig. 2).

On the basis of our observations, we propose that LOY is determined by a ‘common soil’ of shared mechanisms that predispose cells to genomic instability and cancer across many cell types. Our identified *CHEK2* association clearly illustrates this concept: loss of function of *CHEK2* increases LOY in men, and in women delays age at menopause and increases the risk of breast cancer. These effects can all be attributed to inhibited apoptosis in the respective cell types; this is particularly evident in reproductive ageing, for which mouse models demonstrate that *Chek2* is essential for culling oocytes that bear unrepaired DNA double-strand breaks³⁸. The overall trend for alleles that increase LOY to delay age at menopause suggests that many may act through inhibiting the sensing of DNA damage and inhibiting apoptosis, rather than promoting DNA damage (which would lead to earlier menopause owing to premature depletion of the ovarian reserve)^{30,39,40}.

We also note overlap between our identified LOY-associated loci and other complex traits and diseases. For example, seven of the LOY signals that we identify here are correlated with previously reported⁴¹ susceptibility loci for type 2 diabetes (*TP53INP1*, *SUGP1*, *KCNQ1*, *CCND2*, *EIF2S2*, *PTH1R* and *BCL2L1*). At six of these overlapping loci, the allele that confers an increased risk of LOY also increases the risk of type 2 diabetes. *CCND2* encodes cyclin D2, the major D-type cyclin that is expressed in pancreatic β -cells and is essential for adult β cell growth⁴². *TP53INP1* is a p53-inducible gene, the product of which regulates p53-dependent apoptosis. In addition, the LOY-associated genes that encode cyclins and cyclin-dependent kinases (*CCND3*, *CDKN1B* and *CDKN1C*) are also implicated in the growth and maturation of pancreatic β cells. We hypothesize that the previously reported associations between clonal mosaicism in the blood and type 2 diabetes^{15,43} may reflect a common susceptibility to cell-cycle dysregulation and genomic instability, which leads to both increased clonal mosaicism and reduced number of pancreatic β cells. Future studies should aim to more systematically assess the relationships and potential mechanisms that link LOY-associated variants and these broader outcomes on health.

Finally, the ‘common soil’ hypothesis discussed above does not preclude the possibility that LOY in leukocytes also has a direct role in disease, for example, through impaired immune function⁴⁴. A growing awareness of the physiological importance of chromosome Y outside of reproductive development challenges the view of this chromosome as a ‘genetic wasteland’⁴⁵. The male-specific region of the Y chromosome encodes 45 distinct proteins, which have roles in fundamental processes such as chromatin modification (*KDM5D* and *UTY*), gene transcription (*ZFY*) and translation (*DDX3Y*, *EIF1AY* and *RPS4Y1*). Indeed, our observation from scRNA-seq data that leukocytes with LOY have dysregulated expression of autosomal genes supports the notion of a direct physiological effect.

We hope that future experimental studies may build on these observations and will yield further insights into mechanisms that have broad relevance to a range of cancers and other ageing-related diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1765-3>.

- Jacobs, P. A., Brunton, M., Court Brown, W. M., Doll, R. & Goldstein, H. Change of human chromosome count distribution with age: evidence for a sex difference. *Nature* **197**, 1080–1081 (1963).
- Jacobs, P. A., Court Brown, W. M. & Doll, R. Distribution of human chromosome counts in relation to age. *Nature* **191**, 1178–1180 (1961).
- Zhou, W. et al. Mosaic loss of chromosome Y is associated with common variation near *TCL1A*. *Nat. Genet.* **48**, 563–568 (2016).
- Wright, D. J. et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
- Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease — clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
- Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
- Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
- Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
- Vattathil, S. & Scheet, P. Extensive hidden genomic mosaicism revealed in normal tissue. *Am. J. Hum. Genet.* **98**, 571–578 (2016).
- Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
- Forsberg, L. A. et al. Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).
- Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
- Machiela, M. J. et al. Characterization of large structural genetic mosaicism in human autosomes. *Am. J. Hum. Genet.* **96**, 487–497 (2015).
- Dumanski, J. P. et al. Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).
- Loffield, E. et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
- Forsberg, L. A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
- Noveski, P. et al. Loss of Y chromosome in peripheral blood of colorectal and prostate cancer patients. *PLoS One* **11**, e0146264 (2016).
- Machiela, M. J. et al. Mosaic chromosome Y loss and testicular germ cell tumor risk. *J. Hum. Genet.* **62**, 637–640 (2017).
- Ganster, C. et al. New data shed light on Y-loss-related pathogenesis in myelodysplastic syndromes. *Genes Chromosomes Cancer* **54**, 717–724 (2015).
- Loffield, E. et al. Mosaic Y loss is moderately associated with solid tumor risk. *Cancer Res.* **79**, 461–466 (2019).
- Persani, L. et al. Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis. *J. Autoimmun.* **38**, J193–J196 (2012).
- Lleo, A. et al. Y chromosome loss in male patients with primary biliary cirrhosis. *J. Autoimmun.* **41**, 87–91 (2013).
- Grassmann, F. et al. Y chromosome mosaicism is associated with age-related macular degeneration. *Eur. J. Hum. Genet.* **27**, 36–41 (2019).
- Haitjema, S. et al. Loss of Y chromosome in blood is associated with major cardiovascular events during follow-up in men after carotid endarterectomy. *Circ. Cardiovasc. Genet.* **10**, e001544 (2017).
- Dumanski, J. P. et al. Mosaic loss of chromosome Y in blood is associated with Alzheimer disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).
- Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
- Schmidt, M. K. et al. Age- and tumor subtype-specific breast cancer risk estimates for *CHEK2**1100delC carriers. *J. Clin. Oncol.* **34**, 2750–2760 (2016).
- Wang, Z. et al. Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the *TERT*–*CLPTM1L* region on chromosome 5p15.33. *Hum. Mol. Genet.* **23**, 6616–6633 (2014).
- Day, F. R. et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).
- Titus, S. et al. Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, 172ra21 (2013).
- Laine, J., Küntzle, G., Obata, T., Sha, M. & Noguchi, M. The protooncogene *TCL1* is an Akt kinase coactivator. *Mol. Cell* **6**, 395–407 (2000).
- Hirota, T., Gerlich, D., Koch, B., Ellenberg, J. & Peters, J.-M. Distinct functions of condensin I and II in mitotic chromosome assembly. *J. Cell Sci.* **117**, 6435–6445 (2004).
- Petry, S. Mechanisms of mitotic spindle assembly. *Annu. Rev. Biochem.* **85**, 659–683 (2016).
- Godek, K. M., Kabeche, L. & Compton, D. A. Regulation of kinetochore-microtubule attachments through homeostatic control during mitosis. *Nat. Rev. Mol. Cell Biol.* **16**, 57–64 (2015).
- London, N. & Biggins, S. Signalling dynamics in the spindle checkpoint response. *Nat. Rev. Mol. Cell Biol.* **15**, 736–747 (2014).
- Cory, S. & Adams, J. M. The Bcl2 family: regulators of the cellular life-or-death switch. *Nat. Rev. Cancer* **2**, 647–656 (2002).

37. Zaremba, T. et al. Poly(ADP-ribose) polymerase-1 (PARP-1) pharmacogenetics, activity and expression analysis in cancer patients and healthy volunteers. *Biochem. J.* **436**, 671–679 (2011).
38. Bolcun-Filas, E., Rinaldi, V. D., White, M. E. & Schimenti, J. C. Reversal of female infertility by *Chk2* ablation reveals the oocyte DNA damage checkpoint pathway. *Science* **343**, 533–536 (2014).
39. Lin, W., Titus, S., Moy, F., Ginsburg, E. S. & Oktay, K. Ovarian aging in women with *BRCA* germline mutations. *J. Clin. Endocrinol. Metab.* **102**, 3839–3847 (2017).
40. Weinberg-Shukron, A. et al. Essential role of *BRCA2* in ovarian development and function. *N. Engl. J. Med.* **379**, 1042–1049 (2018).
41. Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).
42. He, L. M. et al. Cyclin D2 protein stability is regulated in pancreatic β -cells. *Mol. Endocrinol.* **23**, 1865–1875 (2009).
43. Bonnefond, A. et al. Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
44. Case, L. K. et al. The Y chromosome as a regulatory element shaping immune cell transcriptomes and susceptibility to autoimmune disease. *Genome Res.* **23**, 1474–1485 (2013).
45. Maan, A. A. et al. The Y chromosome: a blueprint for men's health? *Eur. J. Hum. Genet.* **25**, 1181–1188 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

¹Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. ²Department of Genetics, Harvard Medical School, Boston, MA, USA. ³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁴Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁵Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. ⁶Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. ⁷MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK. ⁸Open Targets

Core Genetics, Wellcome Sanger Institute, Hinxton, UK. ⁹Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan. ¹⁰Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. ¹¹Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. ¹²Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan. ¹³deCODE Genetics, Amgen, Reykjavik, Iceland. ¹⁴Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. ¹⁵23andMe, Mountain View, CA, USA. ¹⁶Colon Cancer Genetics Group, Medical Research Council Human Genetics Unit and CRUK Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK. ¹⁷Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. ¹⁸Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. ¹⁹Geriatrics Research Group, Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden. ²⁰Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. ²¹Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²²Division of Molecular Pathology, Institute of Medical Science, University of Tokyo, Tokyo, Japan. ²³Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. ²⁴Faculty of Pharmacy, Medical University of Gdansk, Gdansk, Poland. ²⁵Faculty of Medicine, University of Iceland, Reykjavik, Iceland. ²⁶Cancer Genetics and Evolution Laboratory, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. ²⁷William Harvey Research Institute, Queen Mary University, London, UK. ²⁸School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. ²⁹Lists of participants and their affiliations appear in the Supplementary Information. ³⁰Kyoto-McGill International Collaborative School in Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. ³¹DNRF Center for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. ³²Department of Biochemistry, University of Cambridge, Cambridge, UK. ³³Wellcome Trust and Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. ³⁴Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. ³⁵Beijer Laboratory of Genome Research, Uppsala University, Uppsala, Sweden. ³⁶These authors contributed equally: Lars A. Forsberg, John R. B. Perry. *e-mail: john.perry@mrc-epid.cam.ac.uk

Article

Methods

Data reporting

No statistical methods were used to predetermine sample size as the maximum available sample from the UK Biobank study was utilized. The experiments were not randomized as this approach was not relevant to the study design. The investigators were not blinded to allocation during experiments and outcome assessment as this was not relevant to the study design.

Phenotype preparation in the UK Biobank

We adapted a recently developed statistical approach¹⁰ for detecting autosomal mosaic events to identify male individuals with LOY on the basis of allele-specific genotyping intensities in the PAR of the sex chromosomes. In contrast to previous work, which has quantified Y chromosome loss on the basis of median genotyping intensity over the non-pseudo-autosomal region (non-PAR) of the Y chromosome (mLRR-Y)^{3,4,14,15}, our approach leverages the diploid nature of the PAR to ascertain mosaic Y loss on the basis of differences between maternal (X PAR) and paternal (Y PAR) allelic intensities at heterozygous sites: mosaic Y loss causes Y PAR intensities to decrease relative to X PAR intensities. This intuition can be harnessed even in population cohorts for which absolute phase information (that is, information about maternal versus paternal inheritance of alleles) is unavailable; we can overcome this obstacle by performing statistical phasing and subsequently identifying evidence of an imbalance in allelic intensities between the two statistically phased haplotypes (accounting for the possibility of phase-switch errors)^{9,10}. In general, the signal produced by phased allelic imbalances is typically much cleaner than estimates of total genotyping intensities (for example, mLRR-Y), as the latter can vary substantially across the genome owing to technical artefacts⁴⁶.

We applied this approach to genotyping intensity data from blood DNA from the full UK Biobank cohort (a study described extensively elsewhere⁴⁷), analysing 1,239 genotyped variants on PAR1 that passed quality control (out of 1,301 total PAR1 variants). (We ignored the much-shorter PAR2, which only contained 56 genotyped variants, of which 37 passed quality control.) To maximize phasing accuracy, we phased the full cohort including both males and females using Eagle2⁴⁸, after which we restricted our attention to males. We called mosaic chromosomal alterations (mCAs) in PAR1 using a slightly modified version of the pipeline we described previously¹⁰. Specifically, in our hidden Markov model, we increased the probability of starting in a mosaic state to 0.2; we observed in a preliminary analysis that Y loss events were much more common than autosomal events, so we updated this accordingly in our final analysis to slightly improve the model (and verified that the number of Y loss calls did not drastically change, so no further update was necessary). We also post-processed our PAR1 mCA calls to identify likely mosaic Y loss events on the basis of two criteria: (i) mCA spans the full PAR1 region; and (ii) observed mean \log_2 -transformed R ratio (LRR) is more consistent with a mosaic loss event than a gain or copy number neutral loss of heterozygosity (CNN-LOH) (after taking into account the s.e.m. of LRR and an empirical prior on mCA copy numbers)¹⁰. This procedure produced 44,709 mCA calls in PAR1 (at an estimated false discovery rate (FDR) of 0.05) among 220,924 males who passed sample quality control, of whom 43,306 were classified as likely LOY. These calls contained an average of 321 heterozygous variants on PAR1 that passed quality control and were usually phased perfectly (no switch errors detected by the hidden Markov model in 72% of calls).

We estimated the FDR of PAR-LOY calls using the same phase randomization procedure (similar to permutation testing) that we used previously¹⁰. Specifically, we computed test statistics from PAR-LOY on a batch of control datasets in which we had randomized phase assignments at heterozygous variants (but otherwise kept the data unchanged). The test statistics produced by this procedure gave us an approximate null distribution that we then used to estimate the FDR.

Recalled age at natural menopause was available for 106,237 women with genetic data. We included women who experienced menopause between 40–60 years of age in our analyses, excluding those with menopause that was induced by hysterectomy, bilateral oophorectomy, radiation or chemotherapy, and those who underwent hormone replacement therapy before menopause.

All UK Biobank participants provided written informed consent, the study was approved by the National Research Ethics Service Committee North West—Haydock and all study procedures were performed in accordance with the ethical principles for medical research from the World Medical Association Declaration of Helsinki.

Power comparison of PAR-LOY and mLRR-Y

The efficacy of the PAR-LOY approach relative to mLRR-Y depends primarily on two factors: (i) the relative number of genotyping probes in PAR1 versus Y-nonPAR; and (ii) the size of the cohort (which determines phasing accuracy in PAR1). In the UK Biobank, both factors favour PAR-LOY—the UK Biobank genotyping array contained nearly twice as many PAR1 than Y-non-PAR variants (1,301 versus 691 variants) and the cohort size is extremely large (around 500,000 individuals). For comparison, the BioBank Japan genotyping data contained only one-fifth as many PAR1 compared to Y-non-PAR variants (that is, a relative coverage of PAR1 that is an order of magnitude lower).

To quantify the effects of these two factors on PAR-LOY, we subsampled the UK Biobank dataset to simulate the effects of reduced PAR1 content and reduced phasing accuracy (owing to smaller sample size). Specifically, we applied PAR-LOY to $20 = 4 \times 5$ datasets in which we subsampled (i) the number of PAR1 variants included in the analysis (downsampling once, twice, four and eight times); and (ii) the number of samples included in the analysis (downsampling once, 3.5, 10, 35 and 100 times). In each scenario, we compared the quality of PAR-LOY calls to mLRR-Y by comparing association strength at the top three Y loss GWAS hits (rs17758695 (*BCL2*), rs2887399 (*TCL1A*) and rs59633341 (*TSC22D2*)). We computed relative association strength by taking the mean χ^2 association test statistic across the three variants for PAR-LOY and dividing by the corresponding quantity for mLRR-Y.

We observed (Extended Data Fig. 4) that the PAR-LOY approach benefited considerably from high PAR1 genotyping coverage in the UK Biobank as well as highly accurate phasing (with diminishing returns beyond a cohort size of around 100,000 samples, presumably owing to phasing accuracy becoming near perfect across PAR1). For large cohorts with more than 100,000 samples, our results indicate that the PAR-LOY approach becomes advantageous when a genotyping array contains at least one third (approximately) as many PAR1 as Y-non-PAR variants.

Genetic association testing in the UK Biobank

We used genetic data from the V3 release of UK Biobank⁴⁷, containing the full set of Haplotype Reference Consortium (HRC) and 1000 Genomes imputed variants. In addition to the quality control metrics performed centrally by UK Biobank, we defined a subset of ‘white European’ ancestry samples using a *k*-means-clustering approach that was applied to the first four principal components calculated from genome-wide SNP genotypes. Individuals clustered into this group who self-identified by questionnaire as being of an ancestry other than white European were excluded. After application of quality control criteria, a maximum of 205,011 male participants were available for analysis with genotype and phenotype data. Association testing was performed using a linear mixed model implemented in BOLT-LMM⁴⁹ to account for cryptic population structure and relatedness. Only autosomal genetic variants that were common (minor allele frequency (MAF) > 1%), passed quality control in all 106 batches and were present on both genotyping arrays were included in the genetic relationship matrix. Genotyping chip, age at baseline and ten genetically derived principal components were included as covariates.

We defined statistically independent signals (described as lead or index variants) using clumping at a 1-Mb distance across all imputed variants with $P < 5 \times 10^{-8}$, an imputation quality score > 0.5 and MAF $> 0.1\%$. Genome-wide significant lead variants that shared any correlation with each other as a result of long-range linkage disequilibrium ($r^2 > 0.05$) were excluded from further consideration. These loci were also augmented using approximate conditional analyses implemented in GCTA⁵⁰. For these analyses, secondary signals were only considered if they were uncorrelated ($r^2 < 0.05$) with a previously identified index variant and were significant genome-wide before and after conditional analysis.

The total trait variance of all genotyped SNPs was calculated genome-wide and per chromosome using restricted-estimate maximum likelihood, implemented in BOLT-LMM⁴⁹. The corresponding observed-scale estimate was transformed to the liability scale⁵¹.

Replication

Replication was performed in three independent studies using two separate techniques.

First, we used data generated from the customer base of 23andMe Inc., a consumer genetics company. Genotyping array quality control, imputation and downstream association testing for this study has been described extensively elsewhere⁵². All individuals provided informed consent and answered surveys online according to the human participant protocol of 23andMe, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (<http://www.eandreview.com>). DNA extraction and genotyping were performed on saliva samples by National Genetics Institute (NGI), a Clinical Laboratory Improvement Amendments (CLIA)-licensed clinical laboratory and a subsidiary of Laboratory Corporation of America. Mosaic LOY was estimated by calculating the mLRR-Y (normalized signal intensity) across 274 SNPs on the male-specific region of the Y chromosome (MSY) that are shared and perform well across genotyping platforms, using the protocol described previously⁴. Imputation was performed using a combination of the May 2015 release of the 1000 Genomes Phase 3 haplotypes⁵³ and the imputation reference panel from the UK10K project⁵⁴. Genetic association testing was performed using linear regression in 653,019 male research participants of European ancestry, using age, genetically derived principal components and genotyping platform as covariates. Results were adjusted for a genomic-control inflation factor of 1.129.

Second, we analysed whole-blood genome sequences of 8,715 Icelandic males⁵⁵ (age range 41–105 years; mean 63 years), which had been whole-genome-sequenced by the Illumina method to a mean depth of 37×. As an estimate of chromosome Y copy number, we used the average read depth over chromosome Y (using exclusively X-degenerate regions). This was computed by SAMtools from bam files aligned to hg38 and normalized by genome-wide sequencing coverage for the participant. A total of 12 outlier individuals (copy number greater than 1.25) were excluded. Association analysis was performed using a linear mixed model implemented in BOLT-LMM⁴⁹ (to account for population structure instead of genetic principal components) after inverse normal transformation and adjustment for age at which the blood sample was taken. Effect sizes for \log_2 (chromosome Y copy number) were estimated using robust linear regression (rlm from the R package MASS).

Third, we used a sample of 95,380 Japanese ancestry men from the Biobank Japan project, a study which has been described extensively elsewhere⁵⁶. The study was approved by the ethical committees in the Institute of Medical Science, the University of Tokyo and RIKEN Center for Integrative Medical Science. Mosaic LOY in blood was estimated with the quantitative mLRR-Y measure, using a similar protocol to previous studies⁴. Association testing was performed using a linear mixed model implemented in BOLT-LMM⁴⁹, including age, smoking, disease status (using individual binary covariates to adjust for 35 disease outcomes) and chip array as covariates. Replication test statistics were assessed in

a sensitivity model without the smoking and disease-status covariates to ensure consistency between models.

Genomic feature enrichment

We used a previously modified version of GoShifter^{26,57} to calculate the enrichment of fine-mapped (posterior probability ≥ 0.10) and not fine-mapped (posterior probability < 0.10) variants with genomic annotations by locally shifting the annotations and computing overlaps to approximate the null distribution. Z-scores and odds ratios were calculated from 1,000 permutations, and typical two-tailed P values calculated from the z-score statistic. All annotations were obtained as described previously²⁶.

To identify which tissue types were most relevant to genes involved in LOY, we applied linkage-disequilibrium (LD) score regression⁵⁸ to specifically expressed genes (LDSC-SEG)⁵⁹ and g-chromVAR to haematopoietic accessible chromatin²⁶. For LDSC-SEG, cell-type-specific analyses using GTEx and Epigenome Roadmap annotations were performed using the data available on the LDSC-SEG resource page (<https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses>). For g-chromVAR, haematopoietic-specific analyses were performed using assay for transposase-accessible chromatin using sequencing (ATAC-seq) count matrices as previously processed (https://github.com/caleblareau/singlecell_bloodtraits). g-chromVAR estimates were averaged across 10 different random background sets of peaks. We note that—similar to the derivation of cell-type-specific features or specifically expressed genes in LD score regression analyses—g-chromVAR z-scores represent relative enrichment for specific cell types compared to other input cell types, which allows for discrimination between closely related cell types in the haematopoietic lineage.

Gene-expression integration

We used two approaches to map associated genetic variants to genes via expression effects (eQTLs) in whole blood. First, summary Mendelian randomization (SMR) uses summary-level gene-expression data to map potentially functional genes to trait-associated SNPs⁶⁰. We ran this approach using a meta-analysis of whole-blood eQTL data from 31,684 individuals⁶¹. Only transcripts with no evidence of pleiotropic effects as assessed by the HEIDI metric were considered⁶⁰. Second, we used the recently described transcriptome-wide association study (TWAS) approach⁶² to infer associations via gene expression using three whole-blood datasets (Young Finns Study, Netherlands Twin Registry cohorts and GTEx v.6). All data used are available at <http://gusevlab.org/projects/fusion/>. For all analyses, the significance thresholds were set to adjust for the number of tests performed.

Gene set enrichment analysis

Pathway analysis was performed using two distinct approaches, STRING⁶³ and MAGENTA⁶⁴. For STRING, only the gene closest to one of the 156 lead index variants (maximum distance 500 kb) was included in the analysis. By contrast, MAGENTA performs enrichment analysis using the full genome-wide summary statistic data. In this gene set enrichment analysis (GSEA) approach, each gene in the genome is mapped to a single index SNP with the lowest P value within a 300-kb window. This P value, representing a gene score, is then corrected for confounding factors such as gene size, SNP density and LD-related properties in a regression model. Each mapped gene is then ranked by its adjusted gene score. At a given significance threshold (here the 95th percentile of all gene scores), the observed number of gene scores in a given pathway with a ranked score above the specified threshold percentile is calculated. This observed statistic is then compared to 1,000,000 randomly permuted pathways of identical size. This generates an empirical GSEA P value for each pathway.

We used the Open Targets Platform (<https://www.targetvalidation.org/>) to define gene sets comprising genes involved in cancer susceptibility ($n = 249$), somatic drivers of tumour growth ($n = 394$), targets

of approved or in-trial cancer therapies ($n = 458$), ‘affected pathways’ ($n = 216$ with score = 1) and finally an overall aggregated score for involvement in cancer ($n = 934$ with score = 1). The various data sources, and approach applied by Open Targets to score and prioritize target genes within each of these categories, is described in full at <https://docs.targetvalidation.org/getting-started/scoring>. We arbitrarily defined gene-set membership on the basis of an assigned score greater than 0.8 unless otherwise specified. These pathways were tested for enrichment in downstream analyses using MAGENTA.

Fine-mapping

Regions for fine-mapping were defined by extending 0.5 Mb in both directions from each sentinel and merging when regions overlapped, resulting in 126 total regions. All variants in these regions with $MAF > 0.005$ and imputation quality > 0.6 were fine-mapped. Dosage LD was estimated from the UK Biobank study genotype probability files (.bgen) using 167,020 unrelated white British male individuals (<http://www.nealelab.is/uk-biobank/>). Fine-mapping was then performed using v.1.3 of the FINEMAP software⁶⁵ with default settings allowing for up to five causal variants in each region. The UCSC genome browser was used to view individual variants along with hosted features⁶⁶.

Integration with cancer data and modelling LOY as a causal factor

The NHGRI-EBI GWAS Catalog database was accessed and downloaded on June 25, 2018. The downloaded file was curated to only include studies in which cancer is the associated disease and further filtered to remove variants with association P values greater than 5×10^{-8} . Owing to a potential lag between the time a new GWAS is published and included in the NHGRI-EBI GWAS Catalog, a supplementary literature search of PubMed was performed to identify additional reports of cancer susceptibility studies that were not included in the GWAS Catalog. The literature search was completed on July 18, 2018. LDlink⁶⁷ was used to identify published cancer GWAS-associated genetic variants which are in linkage disequilibrium ($r^2 \geq 0.1$ based on the 1000 Genomes Project European Population data) with one of the 154 LOY lead SNPs. Associations with haematological malignancies were excluded and additional associations were identified by manual searches.

The relationship between LOY-associated variants and cancer was assessed using a series of two-sample summary-statistic-based Mendelian randomization-style analyses. Linear regressions of the cancer log odds ratios (logOR) for each available SNP on the LOY β coefficients were carried out, weighted by the inverse of the variance of the cancer logORs. This is equivalent to an inverse-variance weighted meta-analysis of the variant-specific causal estimates⁶⁸. Because of evidence of overdispersion (that is, heterogeneity in the variant-specific causal estimates), the residual standard error was estimated, making this equivalent to a random-effects meta-analysis. We repeated the analyses using the weighted median method, which is robust to up to half of the genetic variants not being valid instrumental variables⁶⁹. Unbalanced horizontal pleiotropy was tested based on the significance of the intercept term in MR-Egger regression⁷⁰. The analyses were carried out separately for each type of cancer.

Summary statistics for the association between the genetic variants and risk of prostate cancer were obtained from the PRACTICAL and ELLIPSE consortia, based on GWAS analyses of 67,158 prostate cancer cases and 48,350 controls⁷¹. Summary statistics for testicular cancer were obtained from two GWAS studies that were conducted at the Institute of Cancer Research, comprising 4,192 cases of testicular cancer and 12,368 controls^{72,73}. The renal cancer analysis used summary statistics from the kidney cancer GWAS meta-analysis project of 10,784 cases of renal cell carcinoma and 20,406 controls⁷⁴. Colorectal cancer summary statistics were from eight UK-based GWAS studies, totalling 22,372 colorectal cancer cases and 44,271 controls^{75,76}. The summary statistics for overall lung cancer were from GWAS analyses of 29,266

lung cancer cases and 56,450 controls conducted by the International Lung Cancer Consortium⁷⁷. The glioma summary statistics were from GWAS studies of 12,488 cases and 18,169 controls^{78,79}. The breast cancer analysis was based on summary statistics from GWAS analyses of 105,974 breast cancer cases and 122,977 controls conducted by the Breast Cancer Association Consortium (BCAC)⁸⁰, including summary statistics from analyses that were restricted to cases with oestrogen-receptor-positive or oestrogen-receptor-negative breast cancer⁸¹. Summary statistics for the ovarian cancer analysis were from GWAS studies of 25,509 ovarian cancer cases and 48,941 controls conducted by the Ovarian Cancer Association Consortium (OCAC)⁸². The endometrial cancer results were from GWAS studies of 12,906 endometrial cancer cases and 108,979 controls from the Endometrial Cancer Association Consortium (ECAC)⁸³. In addition, analyses for breast and ovarian cancer risk specifically in carriers of a *BRCA1* or a *BRCA2* mutation were carried out using results from GWAS studies conducted by the CIMBA consortium^{81,82}. Although our focus was on the association between LOY and risk of non-haematological cancers, we also included an analysis using summary statistics from GWAS studies of 4,478 chronic lymphocytic leukaemia cases and 13,213 controls⁸⁴. There was some overlap between the control individuals from the breast, ovarian, endometrial and colorectal cancer studies; between the control participants from the prostate, colorectal and testicular cancer studies; and between the control individuals from the endometrial cancer, glioma and chronic lymphocytic leukaemia studies. All of the cancer analyses were based on summary statistics from studies that were restricted to participants with European ancestry.

In addition to the analyses by cancer type, we also used summary statistics from a pan-cancer meta-analysis study of breast, prostate, ovarian and endometrial cancer to look for a more general association between LOY and cancer risk. The pan-cancer summary statistics were derived using a three step procedure. First, the tetrachoric correlation of binary transformed z -scores was used to estimate the correlation between individual cancer summary statistics that is attributable to overlap of control samples⁸⁵. Second, the standard errors from individual cancer summary statistics were decoupled to account for the estimated correlation (resulting from shared controls amongst the meta-analysed cancer strata)⁸⁶; and third, METASOFT software⁸⁷ was used to perform fixed-effect inverse-variance weighted meta-analyses for the combination of four cancers.

Sample preparation for the study of single-cell gene expression

Blood samples from 19 elderly men (median age 80, range 64–89) who were admitted to the Geriatrics Department at Uppsala University Hospital (Uppsala, Sweden) were collected in BD Vacutainer CPT cell-separation tubes containing sodium citrate, and stored on ice. The PBMC fraction was isolated from the whole-blood samples by density gradient centrifugation following the manufacturer’s instructions (BD). PBMCs were collected and suspended in cold $1 \times$ PBS solution with 0.04% BSA. Cell concentrations were measured using an EVE cell counter (NanoEnTek) and diluted to a concentration of 10^6 cells ml^{-1} . All of the prepared samples had a cell viability above 90%. This research was approved by the local research ethics committee in Uppsala, Sweden (Regionala Etikprövningsnämnden EPN; Dnr:2015/092).

Single-cell workflow

We performed scRNA-seq using the 10X Chromium Single Cell 3’ gene-expression solution (10X Genomics) at the SNP&SEQ Technology Platform at Uppsala University (Sweden). This scRNA-seq technology is based on gel beads loaded with barcoded oligonucleotides mixed with single cells and enzymes, before being captured in droplets (gel beads in emulsion; GEMs). The transcripts present in individual cells are barcoded with unique molecular identifiers (UMIs) and used to prepare standard sequencing libraries. All transcripts from single cells get barcoded with the same index sequence, which enables the

transcripts from thousands of single cells to be pooled together in a single sequencing run and allows transcriptional profiling of individual cells. The barcoding and library construction were performed for the 19 PBMC samples using the Chromium Single Cell 3' Reagent Kit (120236/37/62) according to the manufacturer's protocol (CG00052 Single Cell 3' Reagent Kit v.2 User Guide). The entire procedure, from blood sampling to construction of GEMs, was accomplished within 5 h. The generated single-cell libraries were sequenced using a NovaSeq 6000 instrument (Illumina) at the SNP&SEQ Technology Platform and generated a median of 64,900 reads per cell (range 35,213–111,643).

Single-cell bioinformatics pipeline

Sequenced reads were mapped to the human reference GRCh37/hg19 using the software Cell Ranger v.2.0.2 (10X Genomics). Cell Ranger produces a count matrix for each experiment containing the UMI barcodes using sequence information from the 3' end of each transcript in every cell. We used the R package Seurat (v.2.3.1) for further processing and implemented the standard Seurat workflow. Specifically, quality control steps were performed that included the removal of apoptotic cells (that is, cells with more than 5% mitochondrial RNA), as well as the removal of cells with low sequencing coverage and/or a low number of expressed genes (to pass quality control cells had to have at least 350 expressed genes and 800 UMIs). After quality control steps, normalization of the gene expression within each single cell was performed using the function 'NormalizeData' and the built in 'LogNormalize' method of normalization. After this, the most-variable genes were identified using the function 'FindVariableGenes' with the parameters $x.\text{low.cutoff} = 0.2$; $x.\text{high.cutoff} = 4$; $y.\text{cutoff} = 0.5$. All expression values were then scaled using the 'ScaleData' function, principal components were calculated on the basis of the most-variable genes and the number of significant principal components was determined using an elbow plot. Clustering of the dataset was performed using the function 'FindClusters' and cell types for each cluster were determined using canonical marker genes. Refined clustering was achieved by re-clustering within the identified cell types using the above pipeline on subsets of the data. The t -distributed stochastic neighbour embedding (t -SNE) plots were produced using the generated principal components as input.

Determination of LOY in single cells

The LOY status for each sequenced cell was determined on the assumption that cells with LOY would not express genes located in the MSY. Hence, non-LOY blood cells normally express a series of genes located in the MSY (e.g. *EIF1AY*, *RPS4Y1*, *KDM5D* and *ZFY*). We took advantage of this information by scoring LOY in cells without transcripts from the MSY using Ensembl (v.93) to identify all of the genes in the MSY that were included in the data. Cells with a high content of mitochondrial RNA (more than 5%) as well as cells with fewer than 800 UMIs or fewer than 350 genes expressed were removed from the analysis.

Single-cell statistical analyses

To compare differences in autosomal gene expression between LOY cells and non-LOY cells we first performed WilcoxonDETests in the R package Seurat (v.2.3.1). We also developed an in-house random-sampling algorithm to compare gene expression in LOY cells with that in non-LOY cells within specific cell types. First, we established the observed gene expression in LOY cells in the cell type under investigation, by calculating the mean normalized expression values in all participants, within all LOY cells of the tested cell type. Next, we randomly selected, from all participants, a number of cells from the non-LOY cells of the examined cell type, and calculated the mean normalized expression in the resampled cells. To avoid biases caused by inter-individual variation, we programmed the sampling algorithm to sample equal numbers of non-LOY cells and observed LOY cells from each individual. For example, from individuals with 100 LOY cells of a specific cell type, the same number (100) of non-LOY cells from the same cell type was

sampled from the set of non-LOY cells. The resampling of non-LOY cells from all participants was repeated 50,000 times, and for each iteration, the mean normalized expression of the investigated gene in the resampled cells was calculated. The resampled data represent a weighted expression level of the examined gene in non-LOY cells within specific cell types and thus, the resampled distribution represents the normalized expression of the investigated gene in non-LOY cells. The range of variation of gene expression in LOY cells was estimated in a similar fashion, by resampling of a subset of the LOY cells within each individual. Exact P values were calculated by comparing the observed mean expression in LOY cells to the resampled distribution of non-LOY cells. All statistical analyses were performed using R v.3.4.4.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data used in discovery analyses are available from the UK Biobank on request (<https://www.ukbiobank.ac.uk>).

Code availability

All software used in this project is publicly available: BOLT-LMM (v.2.3.2), MoChA (v.1.0), LDSC (v.1.0), MAGENTA (v.2.4), GoShifter (v.1.0), g-chromVAR (v.0.3), SMR (v.0.712), GCTA (v.1.91.6beta), String (v.11.0), FUSION-TWAS (v.1.0), Cell Ranger (v.2.0.2), Seurat (v.2.3.1), FINEMAP (v.1.3) and METASOFT (v.2.0.1).

46. Diskin, S. J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
47. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
48. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
49. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
50. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
51. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
52. Day, F. R. et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* **6**, 8464 (2015).
53. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
54. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
55. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
56. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
57. Trynka, G. et al. Disentangling the effects of colocating genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).
58. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
59. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
60. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
61. Vösa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. Preprint at *bioRxiv* <https://doi.org/10.1101/447367> (2018).
62. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
63. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
64. Segré, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
65. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
66. Casper, J. et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2018).

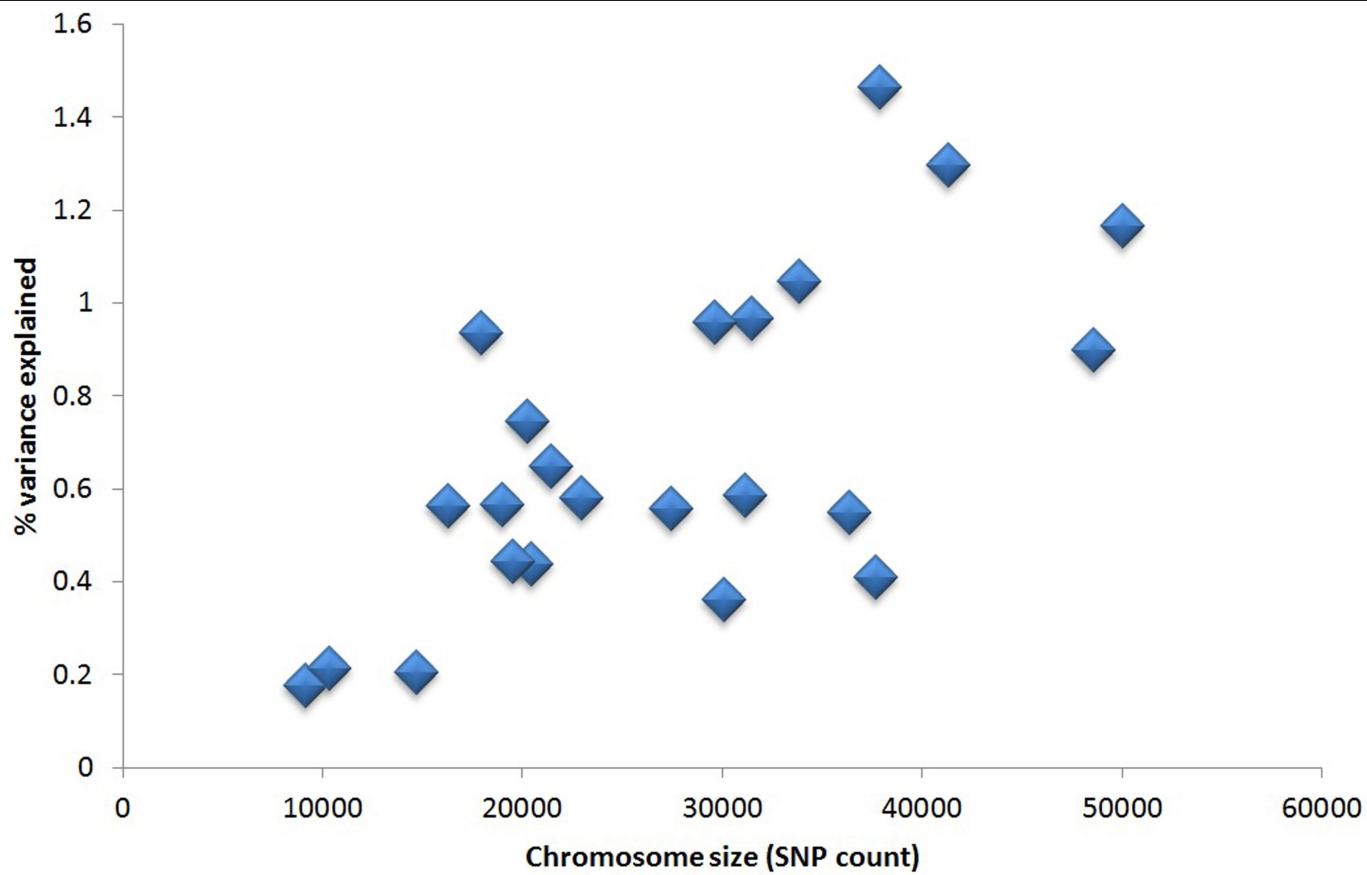
67. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
68. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.* **35**, 1880–1906 (2016).
69. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
70. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
71. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
72. Turnbull, C. et al. Variants near *DMRT1*, *TERT* and *ATF7IP* are associated with testicular germ cell cancer. *Nat. Genet.* **42**, 604–607 (2010).
73. Litchfield, K. et al. Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor. *Nat. Genet.* **49**, 1133–1140 (2017).
74. Scelo, G. et al. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* **8**, 15724 (2017).
75. He, Y. et al. Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: a large Mendelian randomisation study. *BMC Med.* **16**, 142 (2018).
76. May-Wilson, S. et al. Pro-inflammatory fatty acid profile and colorectal cancer risk: a Mendelian randomisation analysis. *Eur. J. Cancer* **84**, 228–238 (2017).
77. McKay, J. D. et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
78. Melin, B. S. et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat. Genet.* **49**, 789–794 (2017).
79. Atkins, I. et al. Transcriptome-wide association study identifies new candidate susceptibility genes for glioma. *Cancer Res.* **79**, 2065–2071 (2019).
80. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
81. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
82. Phelan, C. M. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
83. O'Mara, T. A. et al. Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun.* **9**, 3166 (2018).
84. Law, P. J. et al. Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat. Commun.* **8**, 14175 (2017).
85. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
86. Han, B. et al. A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.* **25**, 1857–1866 (2016).
87. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

Acknowledgements This research was conducted using the UK Biobank Resource under applications 9905 and 19808. The work was supported by the Medical Research Council (unit programme no. MC_UU_12015/2) and the European Research Council (ID no. 679744). J.R.B.P. is grateful to his incredible wife S. Perry, without whose unwavering support his contribution to this work would not be possible. Full study-specific and individual acknowledgements can be found in the Supplementary Information.

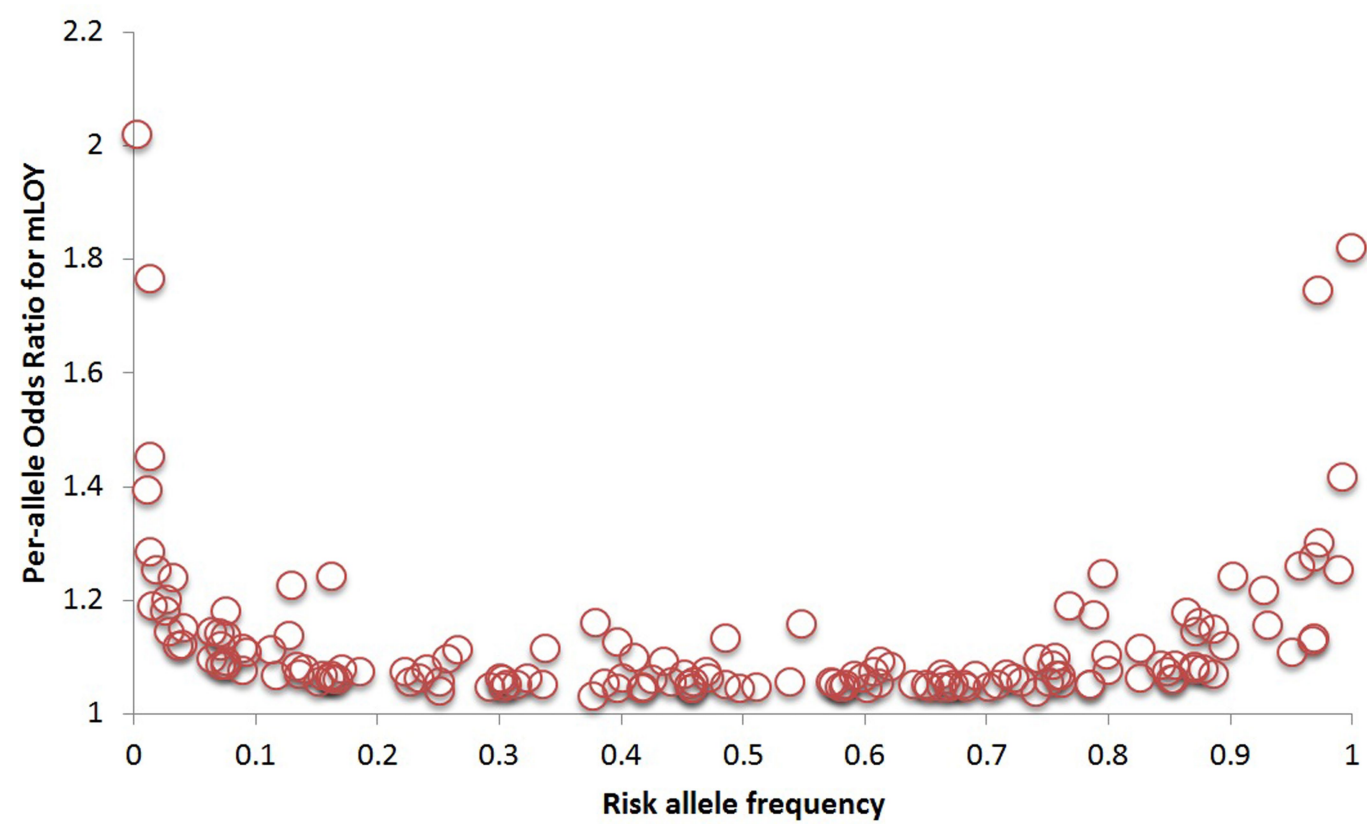
Author contributions All authors reviewed, appraised and edited the manuscript. Primary discovery analyses: D.J.T., J.C.U., D.J.W., F.R.D., J.D., R.A.S., M.J.M., P.-R.L., J.R.B.P.; development of the LOY calling method: G.G., S.A.M., P.-R.L.; design, oversight or analysis of the replication study: C. Terao, O.B.D., P.S., Y.J., F.Z., R.P.K., Y.M., U.T., D.F.G., Y.K., K.S., A.A.; data contribution: M.G.D., D.F.E., V.A.F., R.S.H., S.K., N.D.K., B.K., P.J.L., C.L., I.T., C. Turnbull; scRNA-seq: J.H., M.D., H.D., M.I., J.M., P.O., E.R.-B., B.T.M., J.P.D., L.A.F.; interpretation of findings: D.J.T., R.L., D.F.E., A.M., N.J.W., E.R.H., S.P.J., K.K.O., S.J.C., M.J.M., J.P.D., L.A.F., J.R.B.P.; overall project leadership and first draft writing: J.R.B.P.

Competing interests L.A.F. and J.P.D. are cofounders and shareholders in Cray Innovation AB.

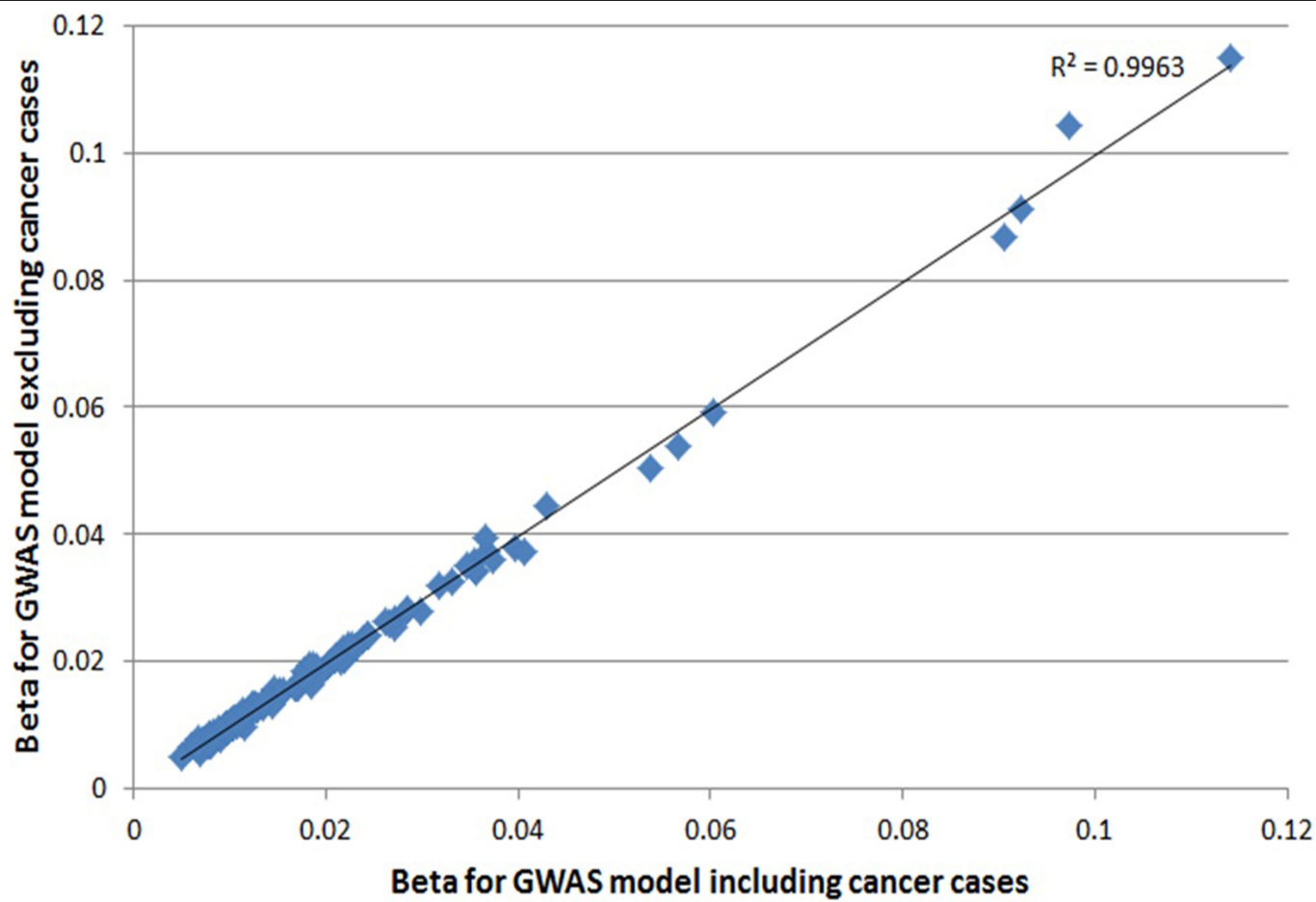
Additional information
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1765-3>.
Correspondence and requests for materials should be addressed to J.R.B.P.
Peer review information *Nature* thanks Don Conrad, Yasminka Jakubek, Paul Scheet and John Witte for their contribution to the peer review of this work.
Reprints and permissions information is available at <http://www.nature.com/reprints>.



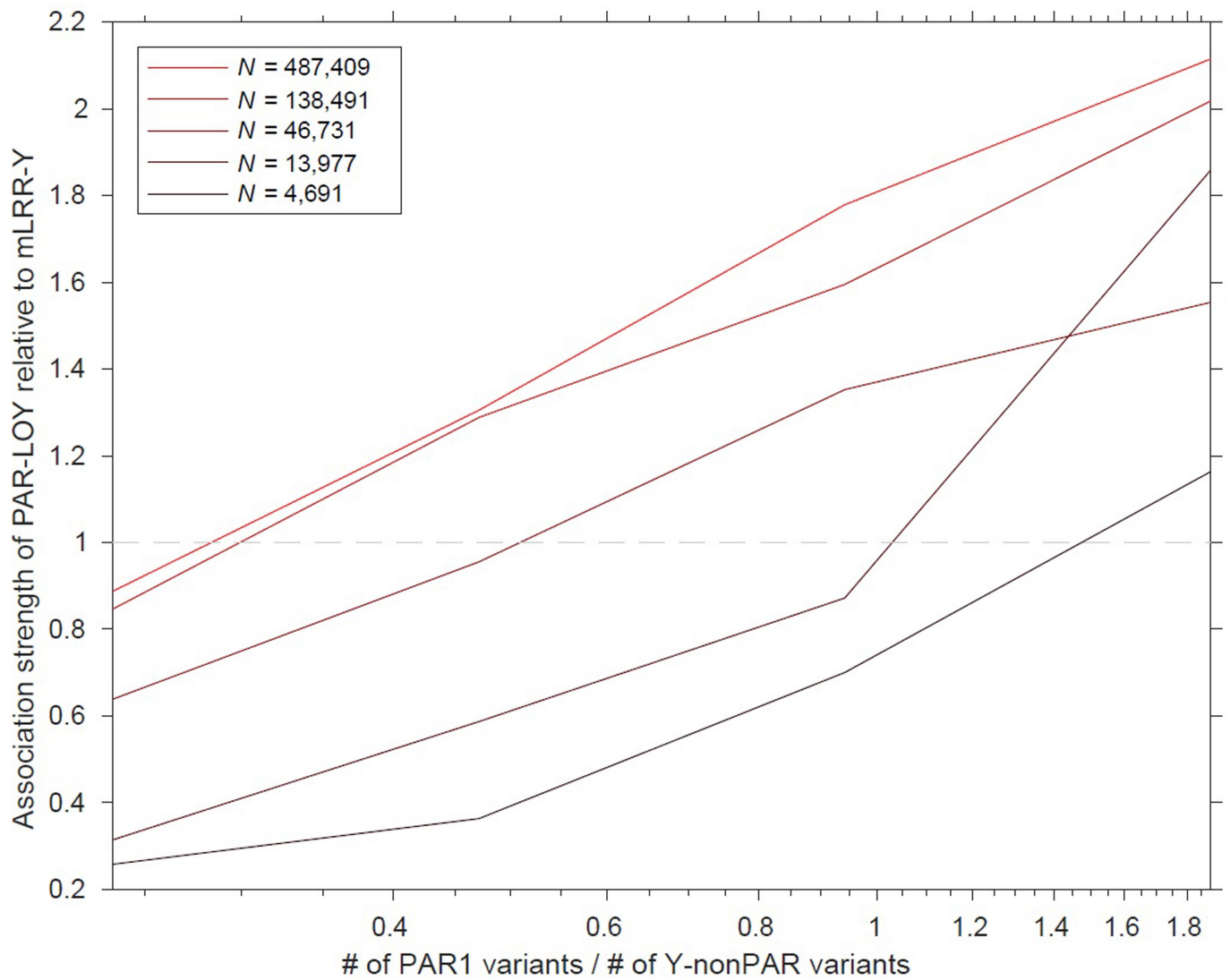
Extended Data Fig. 1 | Liability-scale heritability explained by chromosome. The number of genotyped variants on each chromosome is used as a proxy measure for chromosome size.



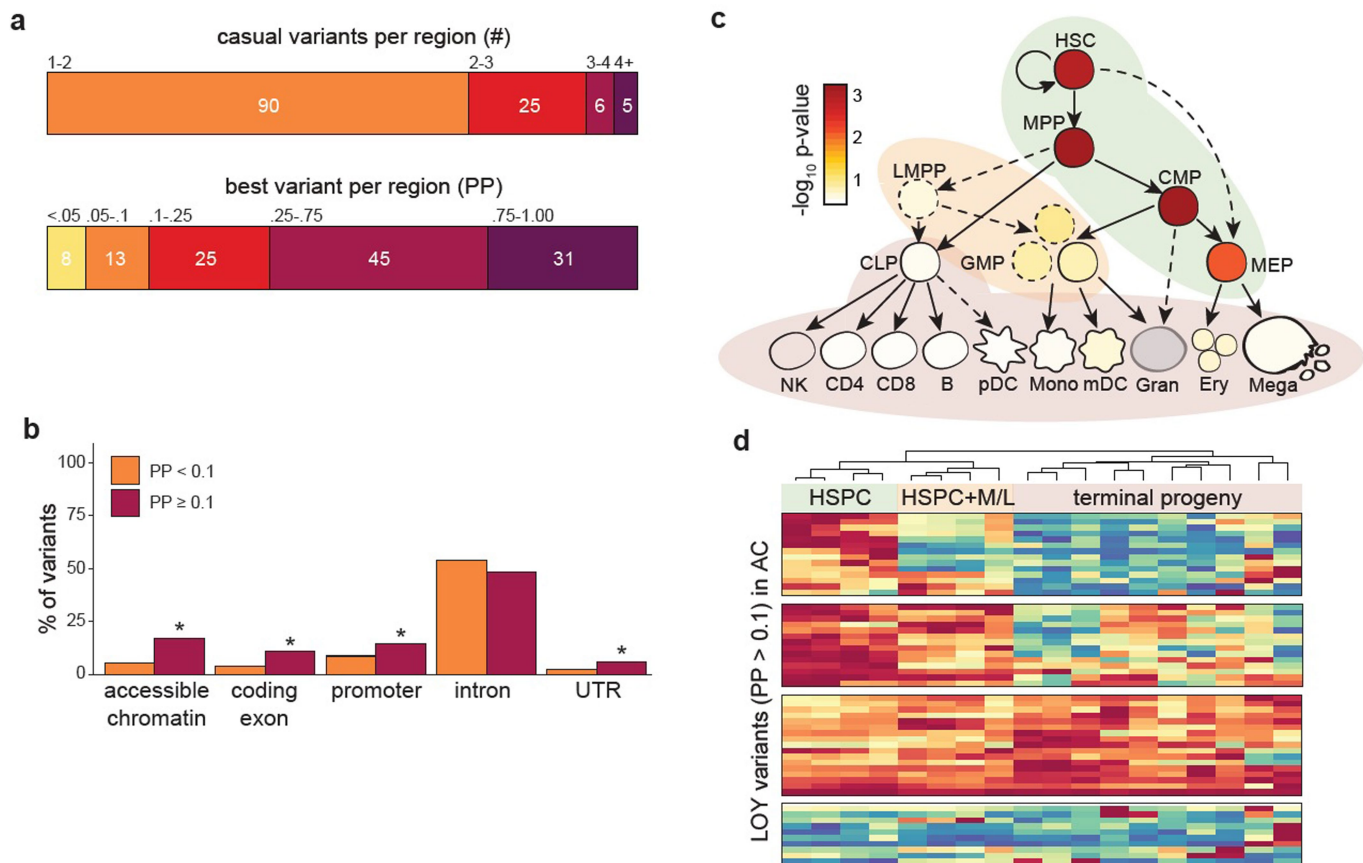
Extended Data Fig. 2 | Distribution of allele frequency and effect size for the 156 identified LOY loci. Estimates of individual SNP effects are taken from the UK Biobank discovery sample.



Extended Data Fig. 3 | Comparison of estimates of SNP β coefficients for the 156 LOY loci in discovery analyses including or excluding cancer cases. Effect estimates were compared between a LOY discovery GWAS analysis either including cancer cases ($n = 205,011$ individuals analysed) or excluding cancer cases ($n = 187,953$ individuals analysed). The squared Pearson correlation coefficient (R^2) is shown.

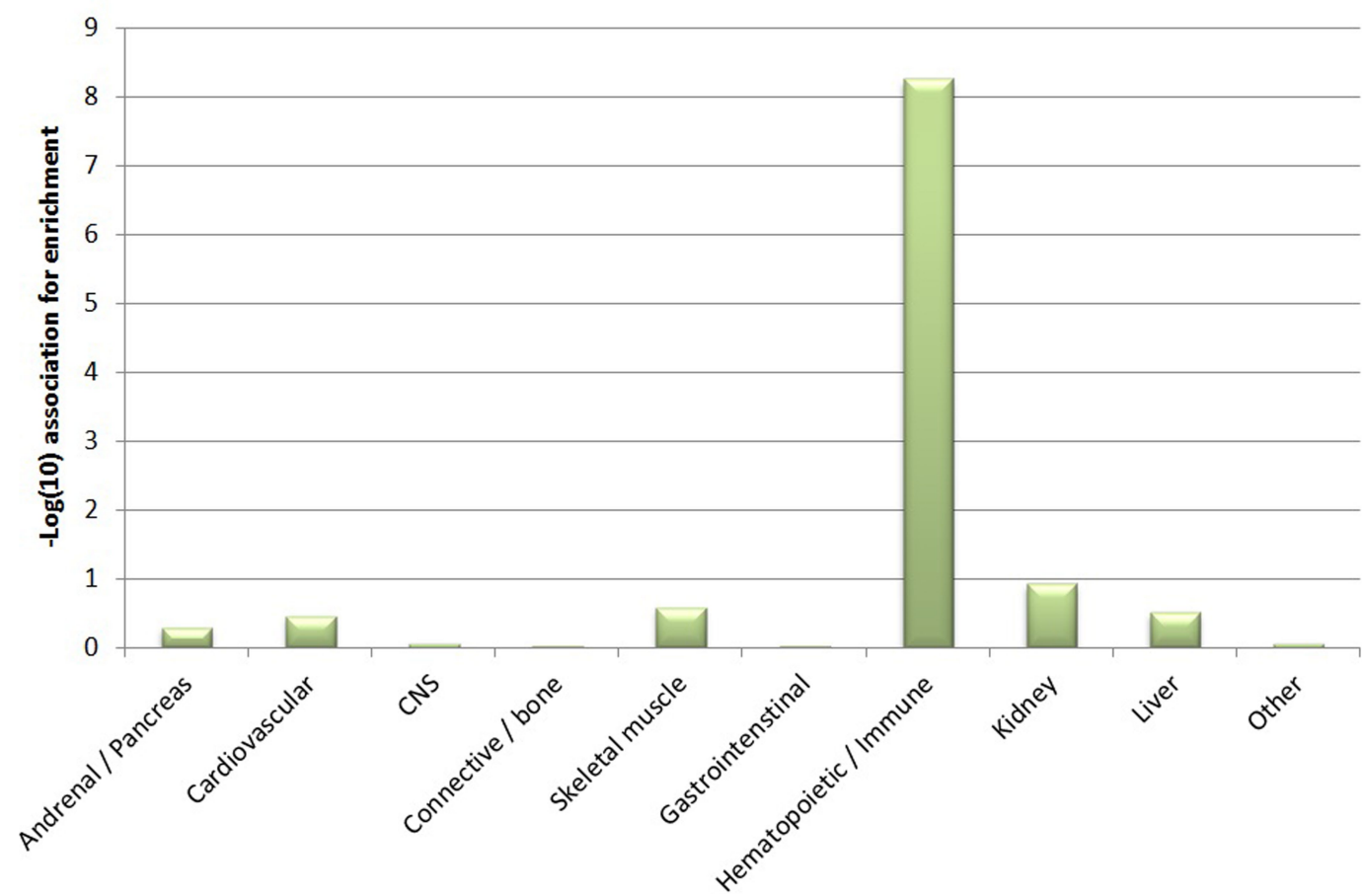


Extended Data Fig. 4 | Power calculation for comparison of LOY calling methods. Plot shows the effect of sample size and the ratio of Y chromosome PAR1 to non-PAR on PAR-LOY power over mLRR-Y.

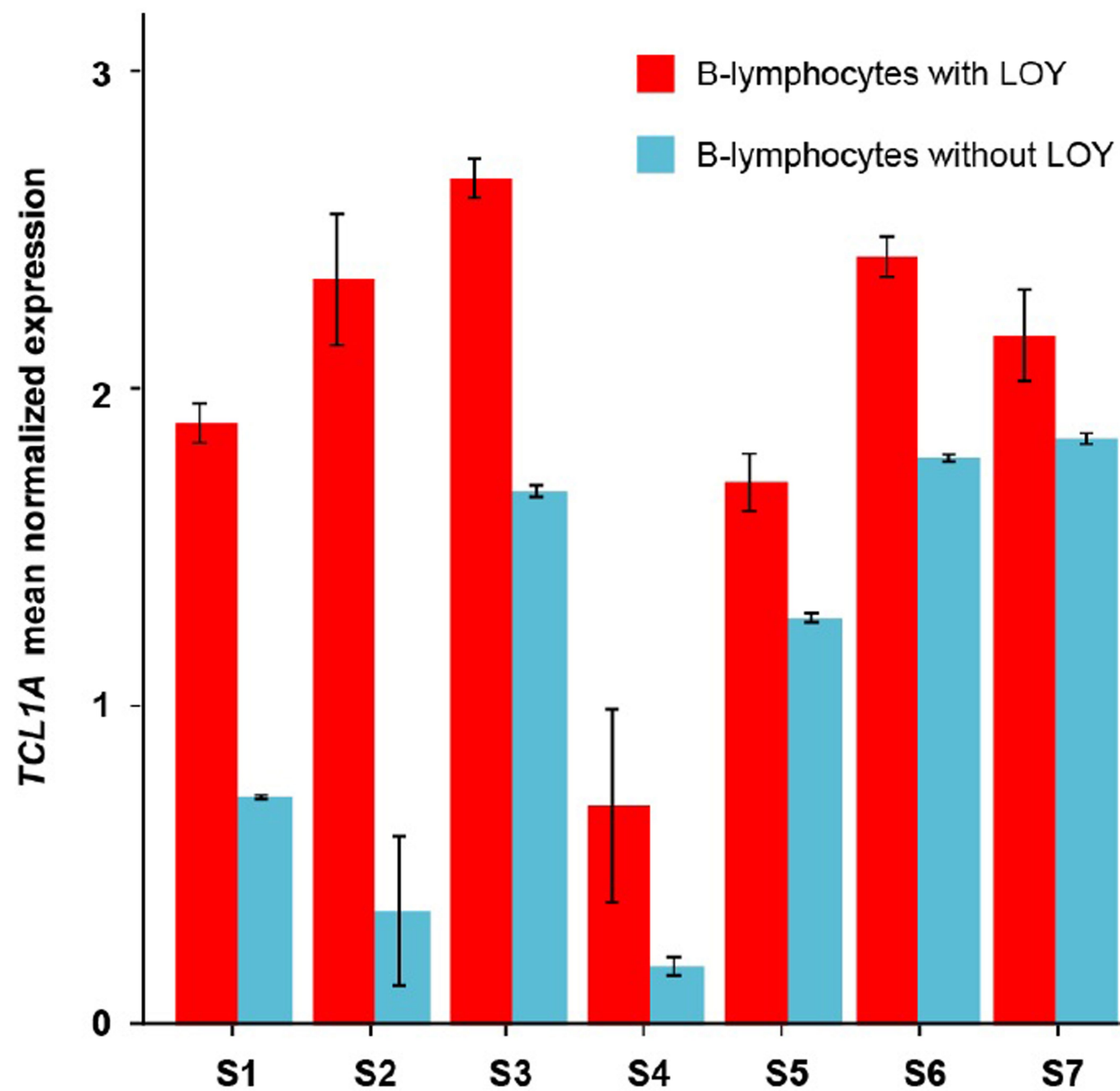


Extended Data Fig. 5 | Results from fine-mapping analyses. a–d. All analyses were performed on genome-wide summary statistic data from the UK Biobank discovery analysis ($n = 205,011$). Two-tailed P values for enrichment were calculated using GoShifter. **a**, The posterior expected number of causal variants (top), as well as the best fine-mapped variant (bottom) in each region. **b**, Genomic enrichments for variants, stratified by posterior probability (PP). Fine-mapped variants were enriched for accessible chromatin in haematopoiesis, as well as in exons, promoters and untranslated regions (UTRs) of protein-coding genes, but not for introns. **c**, Cell-type enrichments (from g-chromVAR analyses) across the haematopoietic tree for LOY. HSCs, multipotent progenitor cells (MPPs) and common myeloid progenitor cells (CMPs) meet the Bonferroni threshold ($\alpha = 0.05/18$). CD4, CD4⁺ T cell; CD8, CD8⁺ T cell; CLP, common lymphoid progenitor cell; ery, erythrocyte; GMP,

granulocyte–macrophage progenitor cell; gran, granulocyte; LMPP, lymphoid-primed multipotent progenitor cell; mDC, myeloid dendritic cell; mega, megakaryocyte; MEP, megakaryocyte–erythroid progenitor cell; mono, monocyte; NK, natural killer cell; pDC, plasmacytoid dendritic cell. **d**, Developmental patterns of accessible chromatin for variants with a posterior probability greater than 10% are shown, revealing that 14 variants are fully restricted to acting within HSPCs, 14 variants can also have regulatory effects in myeloid and lymphoid progenitors, and 17 variants are capable of acting across the majority of haematopoiesis. k -means clustering ($k = 4$ determined by the gap statistic) was used to identify patterns of accessibility, and cell types were hierarchically clustered. AC, accessible chromatin; M/L, myeloid and lymphoid.

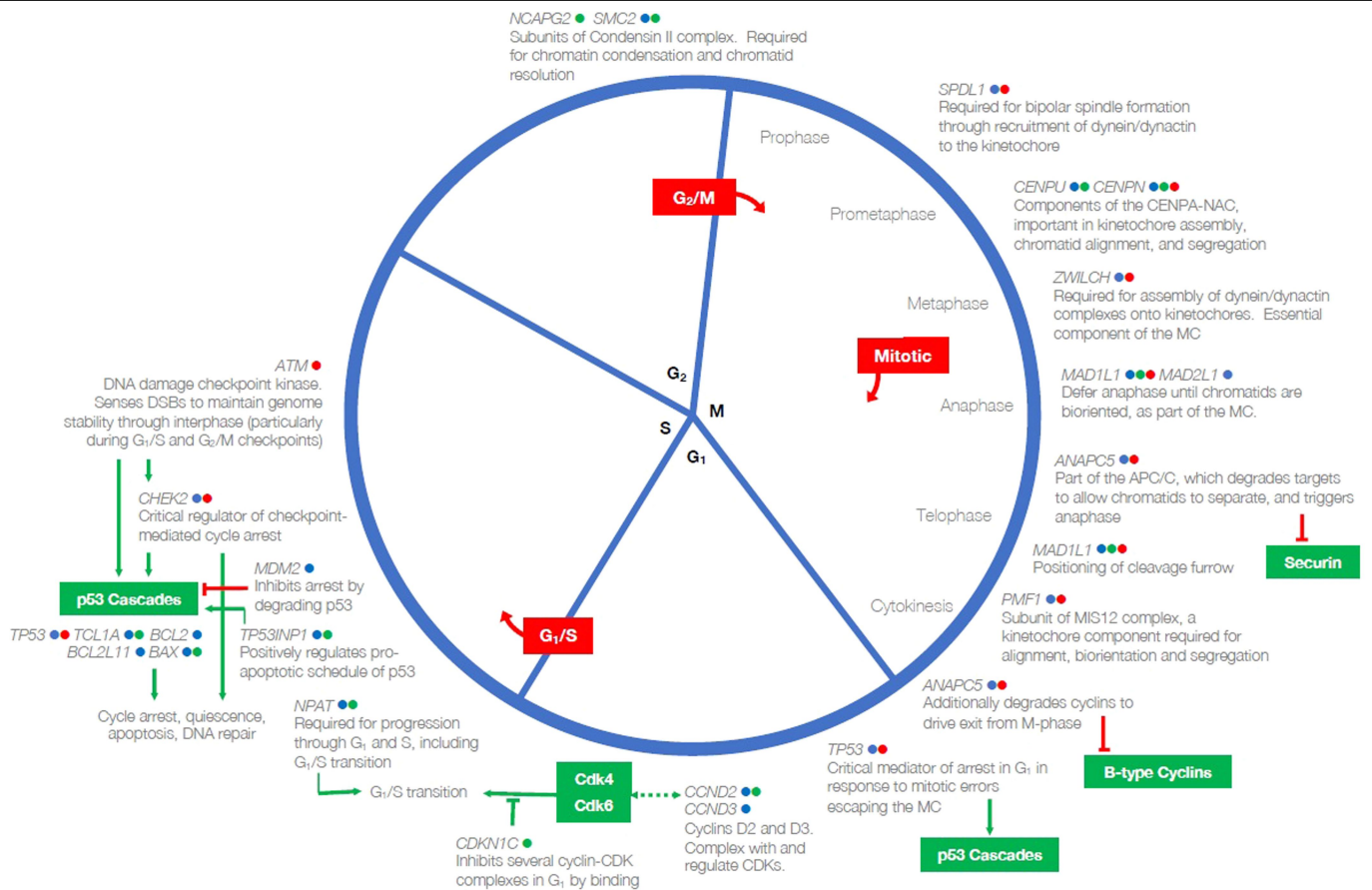


Extended Data Fig. 6 | Estimating cell- and tissue-type enrichment. Analyses performed using LDSC-SEG, with bar chart denoting statistical significance of observed positive enrichment. CNS, central nervous system.



Extended Data Fig. 8 | Differential expression of *TCL1A* in B lymphocytes with and without LOY within individuals. Error bars indicate the 95% confidence interval of the mean normalized expression of *TCL1A* within each group ($n = 277$ B lymphocytes for LOY; $n = 2,459$ for non-LOY). To avoid stochastic effects that might occur in estimations that use a small number of cells, results are shown for individuals with LOY in at least 10% of the B

lymphocytes and with LOY in more than five individual B lymphocytes. Within each of the seven individuals (S1–S7) meeting this criteria, *TCL1A* showed a higher expression in the LOY cells compared to normal cells. This suggests that the observed *TCL1A* overexpression in B lymphocytes without a Y chromosome is independent of the individual genotypes at the lead GWAS SNP (rs2887399).



Extended Data Fig. 9 | Many genes that are associated with LOY converge on mechanistic and regulatory aspects of the cell cycle. All of the genes shown have been prioritized as potentially functional genes at our reported GWAS loci; gene symbols may be shown more than once. Coloured indicators next to each gene symbol specify the type of evidence on which it has been prioritized at its respective locus: blue, nearest protein-coding gene; green, eQTL; red, contains a highly correlated non-synonymous variant. Red boxes indicate each of the three known cell-cycle checkpoints. Red inhibition connectors denote that a target is inhibited by degradation; green that it is inhibited by binding.

Green arrows indicate a signalling cascade and its effector or final physiological effect. Bidirectional dashed green arrows indicate the formation of a complex between the products of the two connected genes. With the exception of p53, proteins contained within green boxes have not been implicated in this GWAS, but are notable interactors of implicated genes. APC/C, anaphase-promoting complex/cyclosome; CDK, cyclin-dependent kinase; CENPA-NAC, CENPA nucleosome-associated complex; MC, mitotic checkpoint.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection N/A - no software was used to collect data for this study

Data analysis BOLT-LMM (v2.3.2), MoChA (v1.0), LDSC (v1.0), MAGENTA (v2.4), GoShifter (v1.0), g-chromVAR (v0.3), SMR (v0.712), GCTA (v1.91.6beta) String (v11.0), FUSION-TWAS (v1.0), Cellranger (v2.0.2), R library Seurat (v2.3.1), FINEMAP (v1.3), METASOFT (v2.0.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in discovery analyses is available from UK Biobank upon request (<https://www.ukbiobank.ac.uk>)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used the full available sample in UK Biobank (N=205,011 men) for discovery analyses. Previously published analyses in~75k men were suitably powered to identify robustly associated loci, so we were confident this expanded sample size was appropriate.
Data exclusions	Only individuals failing standard genotyping quality control parameters defined initially by the UK Biobank study or individuals of non-european ancestry were excluded from analysis. This decision was made prior to performing any downstream analysis.
Replication	We replicated findings in three independent studies (total N=757,114), which included different methods of measuring the trait and an additional ancestry group. All attempted replication has been reported in the manuscript without exception.
Randomization	N/A - randomization occurred naturally as genetic variants were the exposure.
Blinding	N/A - genetic association testing does not require blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	UK Biobank is a national resource that has been described extensively elsewhere (https://www.ukbiobank.ac.uk). Individuals were not directly selected for inclusion in the study on the basis of any disease or health parameter.
Recruitment	UK Biobank: all people aged 40–69 years (men and women) who were registered with the National Health Service and living up to ~25 miles from one of the 22 study assessment centers were invited to participate in 2006–2010. Overall, about 9.2 million invitations were mailed to recruit 503,325 participants (a response rate of 5.47%).
Ethics oversight	UKBB: National Research Ethics Service Committee North West–Haydock and all study procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research. Single cell work: Approved by the local research ethics committee in Uppsala, Sweden (i.e Regionala Etikprovsningsnamnden EPN) Dnr:2015/092.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Chemotaxis as a navigation strategy to boost range expansion

<https://doi.org/10.1038/s41586-019-1733-y>

Received: 12 September 2018

Accepted: 3 October 2019

Published online: 6 November 2019

Jonas Cremer^{1,2,5}, Tomoya Honda^{3,4,5}, Ying Tang¹, Jerome Wong-Ng¹, Massimo Vergassola¹ & Terence Hwa^{1,3*}

Bacterial chemotaxis, the directed movement of cells along gradients of chemoattractants, is among the best-characterized subjects in molecular biology^{1–10}, but much less is known about its physiological roles¹¹. It is commonly seen as a starvation response when nutrients run out, or as an escape response from harmful situations^{12–16}. Here we identify an alternative role of chemotaxis by systematically examining the spatiotemporal dynamics of *Escherichia coli* in soft agar^{12,17,18}. Chemotaxis in nutrient-replete conditions promotes the expansion of bacterial populations into unoccupied territories well before nutrients run out in the current environment. Low levels of chemoattractants act as aroma-like cues in this process, establishing the direction and enhancing the speed of population movement along the self-generated attractant gradients. This process of navigated range expansion spreads faster and yields larger population gains than unguided expansion following the canonical Fisher–Kolmogorov dynamics^{19,20} and is therefore a general strategy to promote population growth in spatially extended, nutrient-replete environments.

Decades of quantitative studies have elucidated how molecular signaling modulates the random run–tumble motion of individual bacterial cells and moves them up chemoattractant gradients^{1–7}. By contrast, the physiological role of chemotaxis remains much less understood¹¹. Notably, many of the chemicals sensed by bacteria are also consumed by them^{17,21}. Hence, cells not only follow the chemoattractants set by their environment, but also shape the spatial profile of attractant abundance and adjust their movement accordingly. In particular, even a small group of cells can form strong attractant gradients that drive cell movement^{9,10,22–24}. Here we perform a quantitative, physiological study of bacterial chemotaxis by taking into account not only chemotaxis itself, but also cell growth and metabolic reactions that lead to the self-generated attractant gradients. We use a motile strain of *E. coli* K-12 for which the growth physiology has been extensively characterized²⁵. As with other motile *E. coli* strains, motility in this strain is enabled by an insertion element that activates the expression of the motility machinery (Supplementary Text 1.1). When inoculated at the centre of a soft-agar plate, these cells swim and readily expand outwards via chemotaxis; migrating cells form a visible ring that propagates with a well-defined speed (Fig. 1a, b), in line with classical observations^{12,17}.

Bacterial expansion dynamics

To characterize this expansion process quantitatively, we first investigated the dependence of expansion speed on the state of cell growth. We used medium containing saturating amounts of a primary carbon source supplemented by small amounts of aspartate or serine as the chemoattractant (see Supplementary Table 1 and Supplementary Text 1 for strains and growth conditions). Growth rate was determined largely

by the primary carbon source, with little contribution from the aspartate and serine supplements (Supplementary Table 2). Expansion speed was clearly affected by the carbon sources used (Extended Data Fig. 1a, Supplementary Table 3). This was not due to the chemotactic effect of these carbon sources, as different expansion speeds were obtained for cells growing in the same medium (glycerol + aspartate, Extended Data Fig. 1b), with different steady-state growth rates attained by titrating the uptake of glycerol²⁶, which is not an attractant²¹ (Supplementary Table 2). The measured expansion speeds follow a common increasing trend with the batch culture growth rate in the respective medium, with either aspartate or serine as the attractant (red and blue symbols, respectively; Fig. 1c). The same relation was also obtained for two other widely studied motile *E. coli* strains (Extended Data Fig. 1c), both of which harbour insertion elements that activate the expression of the motility machinery²⁷ (Supplementary Text 1.1). A more complex medium based on casamino acids that is commonly used in chemotaxis studies supports even faster growth and may be seen as an extension along the same general trend (orange symbols, Fig. 1c). By contrast, the expansion became much slower in the absence of a supplemented attractant, even if the primary carbon source was itself an attractant (for example, with glucose or aspartate only), and for a Δtar mutant that cannot sense the attractant aspartate (Extended Data Fig. 1d).

The positive growth–expansion relation was unexpected in light of the widely held view—which is supported by gene expression data^{14,25,28,29}—that bacterial chemotaxis is specifically triggered by nutrient shortage to find better environments^{14,21,30}. To investigate the origin of this positive relation, we first characterized the swimming speeds of individual cells in different media by recording cell trajectories in well-mixed liquid culture devoid of chemotactic gradients

¹Department of Physics, University of California at San Diego, La Jolla, CA, USA. ²Department of Molecular Immunology and Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands. ³Division of Biological Sciences, University of California at San Diego, La Jolla, CA, USA. ⁴Present address: US Department of Energy, Joint Genome Institute, Berkeley, CA, USA. ⁵These authors contributed equally: Jonas Cremer, Tomoya Honda. *e-mail: hwa@ucsd.edu

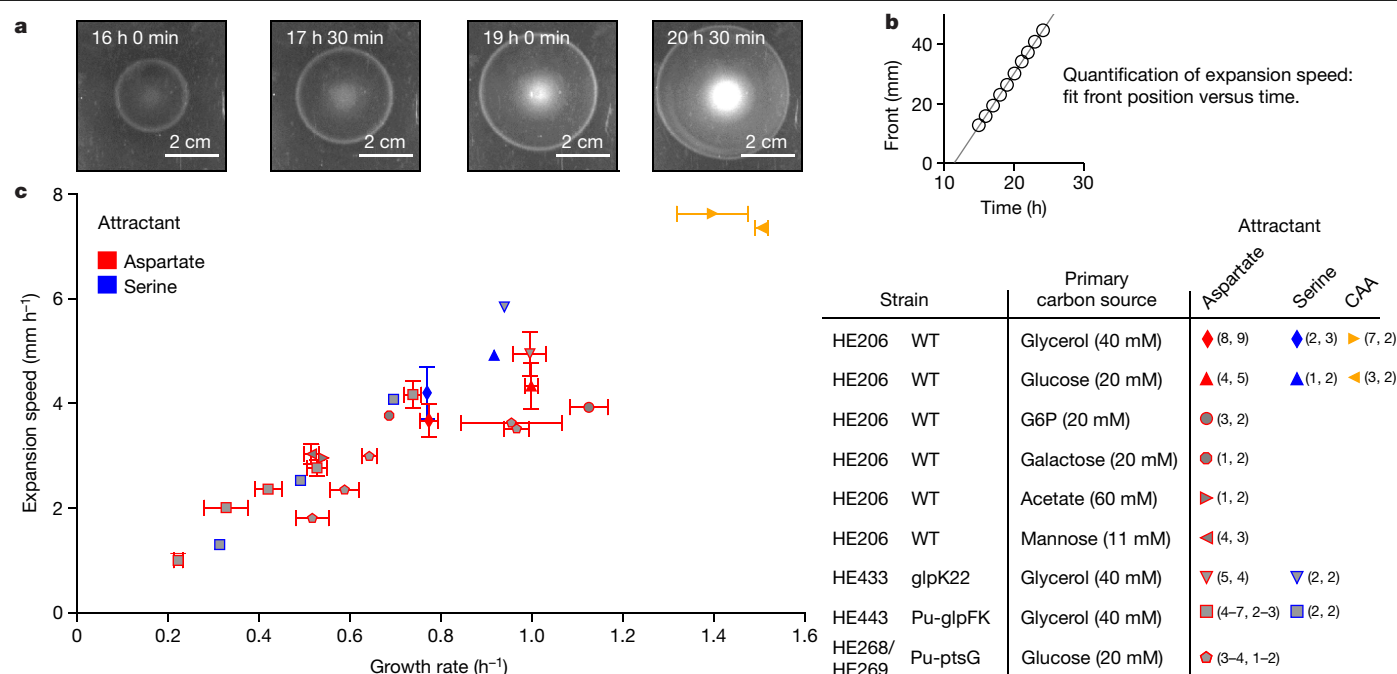


Fig. 1 | Growth dependence of expansion and swimming characteristics. **a**, Population of *E. coli* K-12 HE206 cells (wild-type (WT), see Supplementary Text 1.1) expanding in 0.25% soft agar with 40 mM glycerol and 100 μ M aspartate. Photographs show population density at different times after inoculating exponentially growing cells at the centre of the agar plate. Rings indicate dense bacteria at the population front. Images are representative of experiments repeated independently three times. **b**, Tracking the ring position over time enables the precise quantification of expansion speed (slope). **c**, Expansion speed increases with growth rate. Expansion speeds are shown for HE206 cells (wild-type) in media with different primary carbon sources and for glycerol- and glucose-uptake mutants (HE433, HE443, HE268 and HE269) that

grow at different rates on glycerol or glucose (as controlled by varying inducer levels; Supplementary Text 1.2), in combination with different attractants (100 μ M aspartate, 100 μ M serine or 0.05% casamino acids (CAA)), as indicated. CAA concentration was chosen to have the same aspartate and serine content as the medium with only aspartate or serine. Growth rates were measured in batch culture in the presence of attractant (Supplementary Text 1.4.2, Supplementary Table 2). Notably, these expansion speeds are much larger than those of the Fisher–Kolmogorov dynamics (no more than a few millimetres per hour; see below). Numbers in parentheses indicate number of biological replicates for growth and expansion speed measurements ($n_{\text{growth rate}}$, $n_{\text{expansion speed}}$). Mean \pm s.d. (for $n \geq 3$) are shown.

(Supplementary Text 1.3). Consistent with previous reports^{14,15}, swimming speeds varied strongly in different growth phases (Extended Data Fig. 2a, b). However, such variations resulted from long adaptation periods during outgrowth from starvation, not from a transition out of exponential growth as commonly thought: for all of the *E. coli* strains that we tested, swimming speed remained high throughout steady exponential growth, but decreased rapidly upon entering starvation in both minimal and rich media (Extended Data Fig. 2c, d). Such behaviours, consistent with early findings³¹, suggest that motility is instead favoured by *E. coli* during exponential growth. We next quantified swimming characteristics during exponential growth under different growth conditions, and found that neither swimming speed nor run duration showed substantial variation at different growth rates (Extended Data Fig. 2e, f). It follows that cellular ‘diffusion’ due to random movement resulting from run–tumble dynamics changed little over the broad range of growth rates examined (Extended Data Fig. 2g). Thus, the striking relation between expansion speed and growth rate in Fig. 1c is likely to be a property of the collective dynamics of the propagating population, rather than a direct consequence of single-cell characteristics.

To understand the collective expansion dynamics and its dependence on cell growth, we next observed the spatiotemporal dynamics of fluorescently labelled cells using confocal microscopy at both the population and single-cell levels (Fig. 2a). At the single-cell level, we characterized the random motion of cells in agar by tracking their trajectories over time (Extended Data Fig. 3a, b). The results, quantified by the effective cellular diffusion coefficients across different growth conditions examined, recapitulate the finding from liquid culture that the swimming characteristics are nearly independent of growth rate over the range examined (Extended Data Fig. 3c, orange symbols). At the

population level, we quantified bacterial growth and density profiles over long distances across the entire agar plate. We first established that the growth of bacteria in agar is indistinguishable from that in batch culture (Extended Data Fig. 4a–c). Next, we analysed the time-lapsed radial density profiles of the expanding population in minimal medium, with glycerol as the primary carbon source and a low amount of aspartate as the attractant (Fig. 2b, Supplementary Video 1). The chemotactic ring was seen as a bulge in the density profile following the steep (exponential) increase at the front (Extended Data Fig. 4d). Notably, the advance of the front bulge (approximately 3.2 mm h^{−1}) was steadily followed by a trailing region with a broad, exponentially increasing density profile, suggesting that the outward migration of the ring was tightly coupled to the growth of bacteria behind the ring. This feature was also observed for a population in glycerol supplemented by serine or by aspartate and serine, and in complex medium (Fig. 2c, Extended Data Fig. 4e–h, Supplementary Videos 2–4), suggesting that the underlying dynamical phenomenon is independent of the nature of the attractant. By contrast, for wild-type cells grown in glycerol alone and for Δ tar cells grown in glycerol and aspartate, the much-reduced expansion speeds (Extended Data Fig. 1d) are accompanied by very different (flat) density profiles trailing the steep exponential rise, without recognizable density bulges (see top two panels of Fig. 2d, with full dynamics in Supplementary Video 1 (grey line)). As these populations do not chemotax, their expansion results from a combination of growth rate and random (diffusive) cell movement that can be modelled by the Fisher–Kolmogorov equation^{32,33}. Detailed quantitative analysis established that the bacterial dynamics in these cases was described by the Fisher–Kolmogorov solution without adjustable parameters (Extended Data Fig. 5).

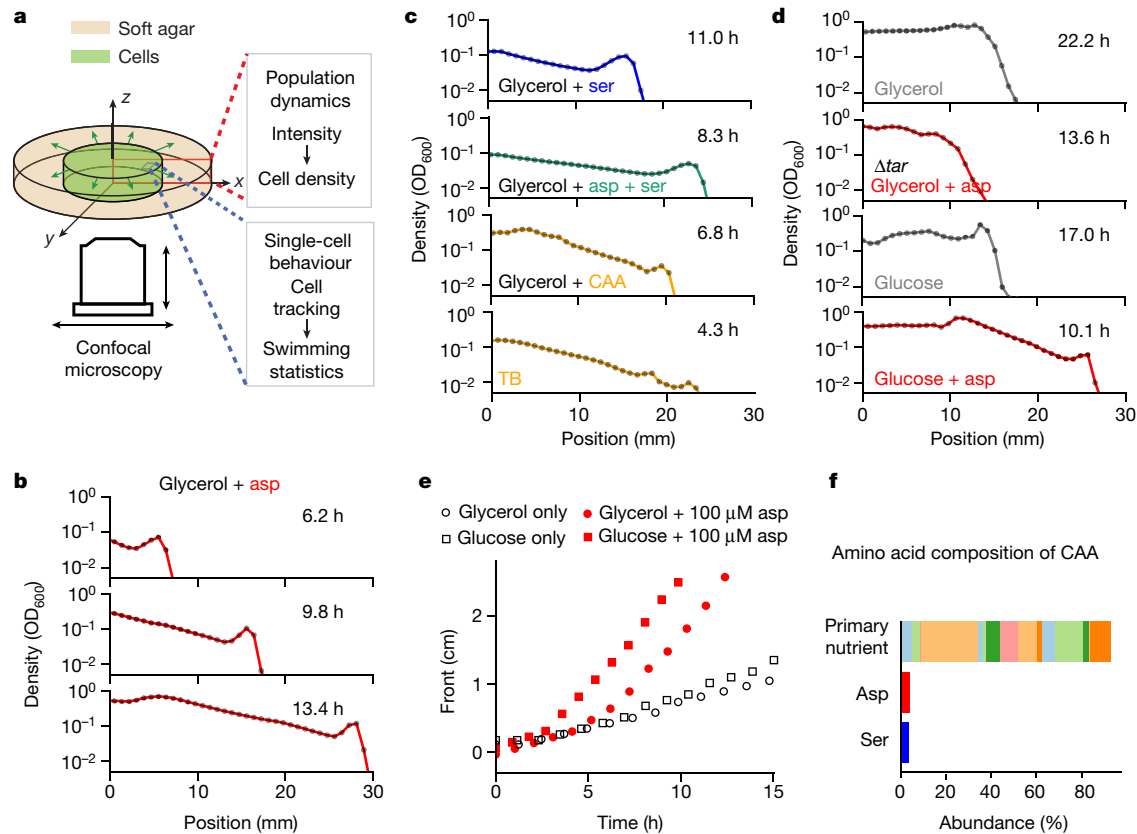


Fig. 2 | Density profiles of expanding bacterial population. **a**, Spatiotemporal evolution of the population in agar was obtained by quantifying the local fluorescence intensity of fluorescently labelled cells throughout the agar using confocal microscopy. Tracking of single cells enabled quantitative characterization of swimming behaviour at various positions and times in agar (Extended Data Fig. 3). **b**, Optical density profiles for fluorescent strain HE274 (wild-type) under reference conditions (40 mM glycerol + 100 μ M aspartate (asp)) at different times, showing an advancing front marked by a density bulge and an exponential trailing region. See Extended Data Fig. 4d for the appearance of the same profiles on a linear density scale. **c**, Single-time density profiles for HE274 in other media with attractant(s), showing the same bulge(s) and trailing exponential region. Time-lapsed density profiles are shown in Extended Data Fig. 4e–h. **d**, No front bulge or trailing exponential region were

seen for wild-type cells in 40 mM glycerol alone or for Δtar cells (HE505) under reference conditions. A bulge without an exponential trailing region was seen for wild-type cells in 20 mM glucose alone, but a front bulge and a trailing exponential region appeared for wild-type cells in 20 mM glucose + 100 μ M aspartate. **e**, Trajectory of front position against time for wild-type cells in 40 mM glycerol or 20 mM glucose, with or without 100 μ M aspartate. Front positions are defined by thresholds in density (optical density at 600 nm (OD_{600}) > 0.002) from confocal scans. **f**, Illustration of our model of a complex medium: most nutrients are lumped together and treated as a ‘primary nutrient’ that fuels cell growth, with small amounts of major chemotactic components (here aspartate and serine). **b–e**, Experiments were conducted twice with similar observations for the reference condition (glycerol + aspartate) and once for other conditions.

Notably, even though glucose is an attractant^{18,21}, cells grown in a glucose plate exhibited a flat density profile following a bulge at the front (Fig. 2d (third panel), Supplementary Video 5), and the population expanded not much faster than in plates with glycerol alone (Fig. 2e, black symbols). Conversely, the combination of glucose and low amounts of aspartate again exhibited a distinct density profile with a broad exponential region trailing the front bulge (Fig. 2d (bottom panel), Extended Data Fig. 4i), along with much faster expansion dynamics (Fig. 2e, red squares). Thus, it appears that the combination of an abundant primary carbon source—regardless of whether it is itself an attractant—supplemented by low amounts of an attractant, is the minimal nutrient requirement to generate the type of behaviour that is generically encountered in rich media (Fig. 2f). Consequently, we adopted the simple medium used in Fig. 2b (glycerol + aspartate) as our model medium (reference condition) in our quantitative study.

The growth–expansion model

To describe the fast expansion dynamics in media with nutrient and attractant, we developed a mathematical model by extending the classical model of a propagating chemotactic ring by Keller and Segel^{34,35}.

The original Keller–Segel front was unstable given a realistic limit of chemotactic sensitivity³⁶. Ingenious models proposed to remedy the problem^{37–39} appear to be too restrictive to capture the simplicity and ubiquity of the observed ring propagation. We focus instead on the crucial role of bacterial growth in driving population expansion. Although the inclusion of growth in chemotactic models also dates back a long time^{40,41}, a satisfactory understanding under general conditions is still lacking^{30,42} (Extended Data Fig. 6). Guided by our experiments, which established the distinct effects of a primary nutrient and a low amount of attractant on expansion (Figs. 1, 2), we explicitly modelled these two ingredients as separate dynamical variables driving bacterial growth and motility; see Fig. 3a for a summary of this growth–expansion (GE) model and Supplementary Text 2 for details.

To test the predictions of the GE model, we determined most model parameters directly by independent experiments in the reference condition, including growth rates in agar (Extended Data Fig. 4c), diffusion coefficients in agar (Extended Data Fig. 3c), and uptake rates of aspartate (Extended Data Fig. 7a–c); Supplementary Table 4 shows the full parameter list. With the molecular parameters available from the literature (Supplementary Text 2.4), only one parameter remained unknown: the chemotactic coefficient χ_0 (Fig. 3a). When we fixed this

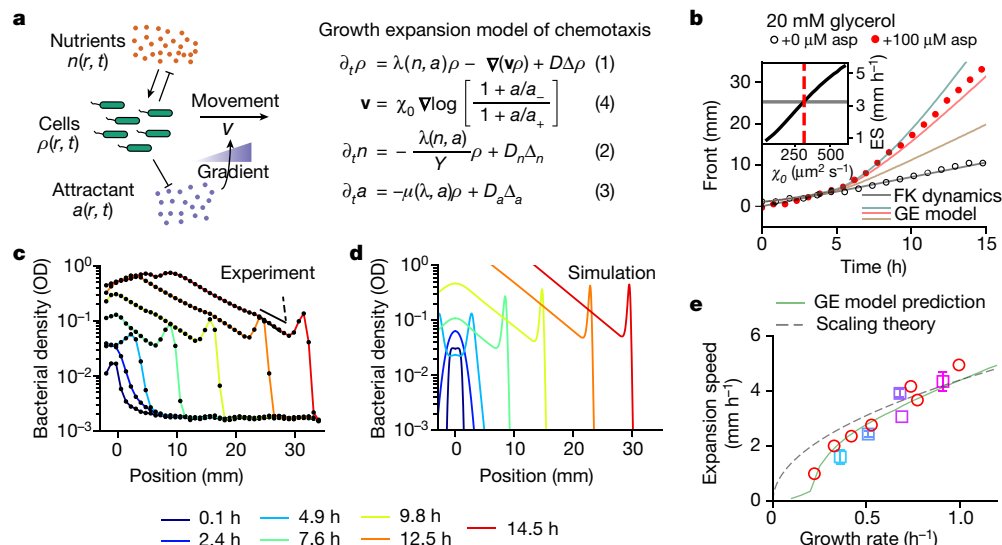


Fig. 3 | The growth–expansion model and its predictions. **a**, The coupled dynamics of growth and migration of the bacterial density, $\rho(r, t)$ is modelled by treating the concentrations of the major nutrient source and the attractant as two distinct variables, $n(r, t)$ and $a(r, t)$, with r being the radial distance from the centre. The dynamics of these variables are given by equations (1)–(3). The rate of cell growth $\lambda(n, a)$ and attractant uptake $\mu(\lambda, a)$ are fixed by our measurements (Supplementary Text 2.1). Nutrients and attractants diffuse with diffusion coefficients D_n and D_a , respectively. Y denotes the growth yield of the nutrient. Following the Keller–Segel model³⁴ and the coarse-grained description of chemotaxis⁴⁸, undirected swimming is described by a cell diffusion term, characterized by D . Directed movement is described by an advection term \mathbf{v} that depends on local attractant gradients (equation (4)), with the chemotactic coefficient χ_0 as the proportionality factor. Other details of the model are described in Supplementary Text 2.1. **b**, The lone unknown parameter of the model χ_0 is fixed by adjusting the ratio χ_0/D such that the expansion speed of the model (black line, inset) matches the experimental observation (horizontal grey line, inset). The corresponding value of χ_0/D (red dashed line, inset) is used for other simulations. Solid red line shows the prediction of the GE model on front position at various times; it captures well the observed dynamics of front propagation of wild-type cells under reference conditions (red circles), including the crossover from Fisher–Kolmogorov

dynamics (black circles and line, Extended Data Fig. 5). Solid green and brown lines show model predictions if χ_0/D was 30% larger or smaller, respectively. **c, d**, Observed (**c**) and simulated (**d**) density profiles at different times (coloured lines). The observed density bulges are less sharp than the simulated ones, owing in part to the limited spatial resolution of the data points (black dots). See Extended Data Fig. 3e for a finer view. Black solid line indicates the predicted slope of the trailing region (Extended Data Fig. 9a); dashed line shows the slope of Fisher–Kolmogorov dynamics for comparison (Extended Data Fig. 5). The fluorescent strain HE274 was used as the wild type in **b, c**; experiments were conducted twice with similar observations. **e**, Predicted and observed changes in expansion speeds when varying the growth rate with different glycerol uptake rate, using HE274, HE484 and HE486 (square symbols). For HE486, three different concentrations of the inducer 3-methylbenzyl alcohol (3MBA) were used. Data for HE274 ($n = 2$) represents a mean of two replicates; expansion speeds for HE484 and HE486 are shown as mean \pm s.d. ($n = 3$ biological replicates) and growth rates from single experiments. Red circles, data from Fig. 1c (glycerol as carbon source). Dashed line, prediction by scaling theory (Extended Data Fig. 9) when changing only the growth rate. Solid green line, prediction of the full GE model, including the observed dependence of model parameters (for example, diffusion constant and uptake rate) on growth rate.

parameter by matching the asymptotic expansion speed of the model with the observed value under the reference condition (Fig. 3b, inset), the GE model quantitatively captured the main features of the population dynamics under the reference condition. The dynamics of the front position (Fig. 3b, red line) captures the data (red circles), including the overlap with Fisher–Kolmogorov dynamics (grey circles and line) at early times. The gross shape of the steady-state density profiles matches well with the experimental profile (Fig. 3c, d); in particular, the slope of the exponential trailing region (Fig. 3c, solid black line) is much flatter than predicted by Fisher–Kolmogorov dynamics (Fig. 3c, dashed black line). A zoomed-in analysis of the front region by using single-cell tracking allowed us to carry out additional comparisons regarding the details of the front region (Extended Data Fig. 3d–h). The model also adequately accounted for the dependence of the expansion speed on the attractant concentration, which has long been known to peak at a moderate level⁴³, for both glycerol and glucose as the primary carbon source (Extended Data Fig. 7d–f). Furthermore, the same model can quantitatively capture the slow expansion dynamics for the case in which the attractant is the sole nutrient (for example, glucose only; Extended Data Fig. 8).

We next applied the GE model to investigate the origin of the positive growth–expansion relation (Fig. 1c). Because the cell diffusion coefficient depends weakly on the growth rate (Extended Data Fig. 3c) whereas the attractant uptake rate yields a strong dependence

(Extended Data Fig. 7c), the latter provides a possible rationalization of the positive growth–expansion relation. Indeed, faster depletion of attractant could naively be thought to allow the front to advance faster. However, this turns out not to be the case, as changing the uptake rate hardly affected the expansion dynamics, in agreement with model predictions (Extended Data Fig. 7g–i, Supplementary Text 2.2). The observed growth–expansion relation (Fig. 1c) can in fact be approximately accounted for within the GE model by just changing the growth rate while keeping all other parameters fixed (Fig. 3e, dashed grey line). Qualitatively, the faster expansion in richer medium results from higher cell density at the front bulge, which leads to faster depletion of the attractant (Extended Data Fig. 7j, k). More quantitatively, a simple scaling analysis captures the square-root form of the growth–rate dependence (Extended Data Fig. 9a–c), and a linear dependence on the chemotactic coefficient χ_0 (Extended Data Fig. 9d, which stands in contrast to the square-root dependence on χ_0 when the attractant is the sole nutrient; Extended Data Fig. 8k). The full model, including the observed growth–rate dependencies of the macroscopic parameters (Extended Data Figs. 3c, 7c) and assuming the independence of the molecular parameters (Supplementary Text 2.1.3), describes the observed data very well (Fig. 3e, green line), without any adjustable parameters. Less is known quantitatively about how serine alone or in combination with aspartate affects the chemotactic parameters. However, the existing data appear to be captured well by the square-root

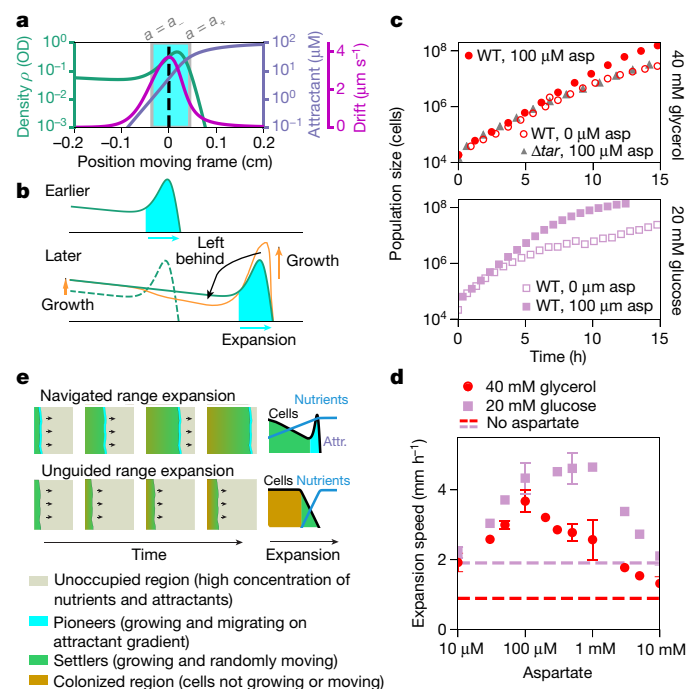


Fig. 4 | The expansion–colonization process. **a**, Spatial profiles of the density (ρ , green), attractant (a , mauve) and drift ($|v|$, purple) during steady migration in co-moving frame; dashed vertical line indicates the position of maximum drift. Highlighted area (cyan) indicates the region where the attractant concentration is in the range between a_- and a_+ (grey lines) in which the chemotactic response is maximal. Full spatiotemporal dynamics of the GE model is shown in Supplementary Video 6 for both the laboratory and co-moving frame. **b**, Illustration of the coupling between the front and trailing regions. Density profiles spaced by one doubling-time are shown. Orange line illustrates the density profile in a hypothetical case in which the effect of diffusion is turned off during this time. The difference (black arrow) represents the effect of cell transfers from the front to the gap right behind the front, which is mediated by diffusion. **c**, Size increase of populations expanding in glycerol or glucose quantified by confocal microscopy (Supplementary Text 1.4). Populations of wild-type cells (coloured symbols, strain HE274) increase faster with than without 100 μM aspartate as attractant. A slower increase is observed for a Δtar mutant that cannot sense aspartate (grey triangles, strain HE505) even if aspartate was present. Differences between the size increases become noticeable after about 6 h, corresponding to the crossover from diffusive Fisher–Kolmogorov dynamics to navigated range expansion (Fig. 3b). Experiments for wild-type cells in glycerol + aspartate were conducted twice with similar results, others once; see also Supplementary Tables 2, 3. **d**, Expansion speed changing with attractant concentration. Wild-type cells (HE206) expanding in glycerol or glucose with varied aspartate concentration (Supplementary Table 9). Expansion speeds without additional attractant are shown as dashed lines. These results confirm previous observations⁴³ and can be quantitatively accounted for by the GE model (Extended Data Fig. 7e, f). Points represent means of $n \geq 2$ biological replicates, error bars (s.d.) shown when $n \geq 3$, see also Supplementary Tables 9, 10. **e**, Navigated mode of range expansion that involves chemotaxis (top) and unguided expansion (Fisher–Kolmogorov dynamics, bottom). Navigation along self-generated gradients of attractants (top) allows faster expansion. Remaining nutrients allow population growth behind the front. Right, corresponding density and nutrient or attractant profiles at the migrating front.

form when analysed separately for each attractant (Extended Data Fig. 1g), suggesting that different attractants and their combinations affect the magnitude but not the functional dependence of population expansion.

Further analysis of the GE model yields insights into how population growth and expansion are generated in a coordinated way. As indicated by the solution of the GE model (Fig. 4a) and confirmed by experimental

observation (Extended Data Fig. 3g), cells in the front bulge region have positive drift velocities and move forward on average; they are the ‘pioneers’ of the population. Conversely, cells behind the bulge experience little chemotactic drift and can grow only locally; they are the ‘settlers’. As the pioneers advance with the front and grow in number, some pioneers would remain behind owing to randomness in cellular motion (described by diffusion of cell density), effectively seeding the void region left behind by the propagating front (black arrow, Fig. 4b) for colonization in the future. This coupled expansion–colonization process is illustrated explicitly in Extended Data Fig. 9e using a discrete agent-based simulation that includes stochastic cell movement and division (see Supplementary Text 3). Cell-to-cell variations in swimming characteristics^{5,24}, which were not included in this calculation for simplicity, are expected to enhance further the transitions between the pioneers and the settlers.

Navigated range expansion

At the population level, the expansion–colonization process sustained by a primary nutrient source and low amounts of attractant provides an effective dispersal mechanism and a clear fitness advantage. This is illustrated by the observed gain in total population size on plates (‘population fitness’) during the expansion process (Fig. 4c). In glycerol medium supplemented with aspartate, the faster propagation of the population front (Extended Data Fig. 1e) accompanies the more rapid increase in total population size compared to a population seeded in the same medium without aspartate (filled and open red circles, respectively, Fig. 4c, top). This fitness advantage for the population is not due to the addition of aspartate as a nutrient supplement, because a Δtar mutant that cannot respond to aspartate chemotactically (Fig. 2d) does not gain any fitness advantage from an aspartate supplement (grey triangles). Furthermore, the aspartate supplement substantially increases the population fitness even if the primary nutrient is itself an attractant such as glucose (purple squares, Fig. 4c, bottom). Thus, the fitness gain specifically requires an environment in which an attractant supplements an abundant primary nutrient source, regardless of whether the nutrient source is an attractant itself, reflecting the requirement for attaining a boost in expansion speed as established earlier (Extended Data Fig. 1e). This advantage of chemotaxis further relies on sufficiently low concentrations of the supplemented attractant (Fig. 4d). Notably, this requirement is at odds with the notion of cells seeking the attractant as a source of nutrients for growth and suggests instead that the supplemented attractant acts as a signaling cue to navigate and accelerate population expansion. Indeed, by separating the roles of nutrients and navigation cues, *E. coli* can use metabolites such as aspartate and serine that are rapidly taken up^{44,45} as strong attractants without concerns for their poor ability to support growth (Supplementary Table 2), while maintaining fast growth on better nutrient sources such as glucose and glycerol.

In summary, chemotaxis along self-generated gradients of low-dose attractant supplements provides a local ‘guide’ for populations to expand rapidly into unoccupied territories, thereby giving them strong fitness advantages to grow in nutrient-replete environments. This navigated mode of range expansion (Fig. 4e, top) contrasts with the canonical, unguided mode of range expansion (Fisher–Kolmogorov dynamics^{19,20,46}, Extended Data Fig. 5) in which the population advances through the growth and random motion of cells at the front, leaving no nutrients behind the front (Fig. 4e, bottom). Notably, in navigated expansion, the guide is provided well before the population experiences any starvation. It thus manifests a built-in diversification strategy for a population with the ‘foresight’ to conquer new territories well before nutrients are depleted in the current environment. This foresight is important, because when starvation is experienced it is likely to be too late to turn on cell motility in order to facilitate effective population expansion (Extended Data Fig. 10).

The strategy of navigated range expansion in nutrient-replete conditions, analysed here for *E. coli* K-12 strains, is one of a number of ways in which chemotaxis can contribute to bacterial fitness. Under other conditions, chemotaxis can be used to respond to starvation or to escape from harsh environments such as non-optimal temperatures or pH ranges^{14,15,21,30}. On the other hand, navigated range expansion, in which a diversified population of pioneers and settlers enables rapid occupation of open habitats for future colonization, might be used for range expansion by organisms other than bacteria. To efficiently guide the movement of the population along the desired direction for expansion, all that is needed is a component of the unoccupied environment that can be easily sensed and modified (for example, degraded but not consumed⁴⁷); this is much easier for higher organisms to accomplish. Thus, navigated range expansion might also be used by higher organisms to rapidly colonize spatially extended habitats.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1733-y>.

- Berg, H. C. *E. coli in Motion* (Springer, 2004).
- Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
- Sourjik, V. & Berg, H. C. Functional interactions between receptors in bacterial chemotaxis. *Nature* **428**, 437–441 (2004).
- Bray, D. & Duke, T. Conformational spread: the propagation of allosteric states in large multiprotein complexes. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 53–73 (2004).
- Korobkova, E., Emonet, T., Vilar, J. M. G., Shimizu, T. S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **428**, 574–578 (2004).
- Tu, Y., Shimizu, T. S. & Berg, H. C. Modeling the chemotactic response of *Escherichia coli* to time-varying stimuli. *Proc. Natl Acad. Sci. USA* **105**, 14855–14860 (2008).
- Sourjik, V. & Wingreen, N. S. Responding to chemical gradients: bacterial chemotaxis. *Curr. Opin. Cell Biol.* **24**, 262–268 (2012).
- Tu, Y. Quantitative modeling of bacterial chemotaxis: signal amplification and accurate adaptation. *Annu. Rev. Biophys.* **42**, 337–359 (2013).
- Waite, A. J. et al. Non-genetic diversity modulates population performance. *Mol. Syst. Biol.* **12**, 895 (2016).
- Baym, M. et al. Spatiotemporal microbial evolution on antibiotic landscapes. *Science* **353**, 1147–1151 (2016).
- Hein, A. M., Carrara, F., Brumley, D. R., Stocker, R. & Levin, S. A. Natural search algorithms as a bridge between organisms, evolution, and ecology. *Proc. Natl Acad. Sci. USA* **113**, 9413–9420 (2016).
- Adler, J. Chemoreceptors in bacteria. *Science* **166**, 1588–1597 (1969).
- Maeda, K., Imae, Y., Shioi, J. I. & Oosawa, F. Effect of temperature on motility and chemotaxis of *Escherichia coli*. *J. Bacteriol.* **127**, 1039–1046 (1976).
- Amsler, C. D., Cho, M. & Matsumura, P. Multiple factors underlying the maximum motility of *Escherichia coli* as cultures enter post-exponential growth. *J. Bacteriol.* **175**, 6238–6244 (1993).
- Staropoli, J. F. & Alon, U. Computerized analysis of chemotaxis at different stages of bacterial growth. *Biophys. J.* **78**, 513–519 (2000).
- Paulick, A. et al. Mechanism of bidirectional thermotaxis in *Escherichia coli*. *eLife* **6**, e26607 (2017).
- Adler, J. Chemotaxis in bacteria. *Science* **153**, 708–716 (1966).
- Koster, D. A., Mayo, A., Bren, A. & Alon, U. Surface growth of a motile bacterial population resembles growth in a chemostat. *J. Mol. Biol.* **424**, 180–191 (2012).
- Skellam, J. G. Random dispersal in theoretical populations. *Biometrika* **38**, 196–218 (1951).
- Hastings, A. et al. The spatial spread of invasions: new developments in theory and evidence. *Ecol. Lett.* **8**, 91–101 (2005).
- Adler, J., Hazelbauer, G. L. & Dahl, M. M. Chemotaxis toward sugars in *Escherichia coli*. *J. Bacteriol.* **115**, 824–847 (1973).
- Saragosti, J. et al. Directional persistence of chemotactic bacteria in a traveling concentration wave. *Proc. Natl Acad. Sci. USA* **108**, 16235–16240 (2011).
- Wong-Ng, J., Melbinger, A., Celani, A. & Vergassola, M. The role of adaptation in bacterial speed races. *PLOS Comput. Biol.* **12**, e1004974 (2016).
- Fu, X. et al. Spatial self-organization resolves conflicts between individuality and collective migration. *Nat. Commun.* **9**, 2177 (2018).
- Hui, S. et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol. Syst. Biol.* **11**, 784 (2015).
- You, C. et al. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature* **500**, 301–306 (2013).
- Barker, C. S., Prüss, B. M. & Matsumura, P. Increased motility of *Escherichia coli* by insertion sequence element integration into the regulatory region of the *flhD* operon. *J. Bacteriol.* **186**, 7529–7537 (2004).
- Liu, M. et al. Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J. Biol. Chem.* **280**, 15921–15927 (2005).
- Zhao, K., Liu, M. & Burgess, R. R. Adaptation in bacterial flagellar and motility systems: from regulon members to ‘foraging’-like behavior in *E. coli*. *Nucleic Acids Res.* **35**, 4441–4452 (2007).
- Lauffenburger, D., Kennedy, C. R. & Aris, R. Traveling bands of chemotactic bacteria in the context of population growth. *Bull. Math. Biol.* **46**, 19–40 (1984).
- Adler, J. & Templeton, B. The effect of environmental conditions on the motility of *Escherichia coli*. *J. Gen. Microbiol.* **46**, 175–184 (1967).
- Fisher, R. The wave of advance of advantageous genes. *Ann. Eugen.* **7**, 355–369 (1937).
- Kolmogorov, A., Petrovsky, I. & Piskounov, N. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Mosk. Univ. Bull. Math.* **1**, 37 (1937).
- Keller, E. F. & Segel, L. A. Model for chemotaxis. *J. Theor. Biol.* **30**, 225–234 (1971).
- Keller, E. F. & Segel, L. A. Traveling bands of chemotactic bacteria: a theoretical analysis. *J. Theor. Biol.* **30**, 235–248 (1971).
- Novick-Cohen, A. & Segel, L. A. A gradually slowing travelling band of chemotactic bacteria. *J. Math. Biol.* **19**, 125–132 (1984).
- Budrene, E. O. & Berg, H. C. Dynamics of formation of symmetrical patterns by chemotactic bacteria. *Nature* **376**, 49–53 (1995).
- Brenner, M. P., Levitov, L. S. & Budrene, E. O. Physical mechanisms for chemotactic pattern formation by bacteria. *Biophys. J.* **74**, 1677–1693 (1998).
- Saragosti, J. et al. Mathematical description of bacterial traveling pulses. *PLOS Comput. Biol.* **6**, e1000890 (2010).
- Nossal, R. Growth and movement of rings of chemotactic bacteria. *Exp. Cell Res.* **75**, 138–142 (1972).
- Lapidus, I. R. & Schiller, R. A model for traveling bands of chemotactic bacteria. *Biophys. J.* **22**, 1–13 (1978).
- Tindall, M. J., Maini, P. K., Porter, S. L. & Armitage, J. P. Overview of mathematical approaches used to model bacterial chemotaxis II: bacterial populations. *Bull. Math. Biol.* **70**, 1570–1607 (2008).
- Wolfe, A. J. & Berg, H. C. Migration of bacteria in semisolid agar. *Proc. Natl Acad. Sci. USA* **86**, 6973–6977 (1989).
- Prüss, B. M., Nelms, J. M., Park, C. & Wolfe, A. J. Mutations in NADH:ubiquinone oxidoreductase of *Escherichia coli* affect growth on mixed amino acids. *J. Bacteriol.* **176**, 2143–2150 (1994).
- Yang, Y. et al. Relation between chemotaxis and consumption of amino acids in bacteria. *Mol. Microbiol.* **96**, 1272–1282 (2015).
- Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl Acad. Sci. USA* **104**, 19926–19930 (2007).
- Seymour, J. R., Simó, R., Ahmed, T. & Stocker, R. Chemoattraction to dimethylsulfoniopropionate throughout the marine microbial food web. *Science* **329**, 342–345 (2010).
- Celani, A. & Vergassola, M. Bacterial strategies for chemotaxis response. *Proc. Natl Acad. Sci. USA* **107**, 1391–1396 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Strains used in this study

The reference strain for this study is HE206, a motile variant of an *E. coli* K-12 strain NCM3722 for which the physiology has been well characterized^{25,26,49–52}. Similar to other motile *E. coli* strains that have previously been studied^{27,53}, the strain carries an insertion element upstream of the *flhDC* operon to enable motility. See Supplementary Text 1.1 for the strain context. Details on all used strains (deletion mutants, titratable carbon-uptake strains and fluorescently labelled strains) and their construction are provided. As indicated in the text, comparisons are made with MG1655 and RP437, other commonly used motile *E. coli* strains. All strains used in this study are listed in Supplementary Table 1.

Growth media

All growth media used in this study were based on a modified MOPS-buffered minimal medium⁵⁴. Trace micronutrients were not added into the MOPS medium, as the metal components have been reported to inhibit motility of *E. coli*³¹. To change growth conditions, different carbon sources were supplemented in the medium. When indicated, CAA and tryptone broth (TB) were used. Minimal medium for the growth of RP437 involves four additional amino acids. For the strains with titratable carbon uptake (glycerol or glucose), 3-MBA was additionally provided as the inducer. Full details on medium composition and concentrations are provided in Supplementary Text 1.2.

Strain culturing and growth rate measurement

Growth measurements were performed in a 37 °C water bath shaker operating at 250 rpm. The culture volume was no more than 4.5 ml in 18 mm × 150 mm test tubes (Fisher Scientific) to limit the depth of the culture in the tube for aeration purposes. Each growth experiment was carried out in three steps: seed culture in LB broth, pre-culture, and experimental culture in identical minimal medium. For the seed culture, one colony from a fresh LB agar plate was inoculated into liquid LB broth and cultured at 37 °C with shaking. After 4–5 h, cells were centrifuged and washed once with the desired minimal medium. Cells were then diluted into the minimal medium and cultured in a 37 °C water bath shaker overnight (pre-culture). The starting OD₆₀₀ in pre-culture was adjusted so that exponential cell growth was maintained overnight, preventing cells from reaching saturation. Cells from the overnight pre-culture were then diluted to OD₆₀₀ = 0.005–0.02 in identical pre-warmed minimal medium, and cultured in a 37 °C water bath shaker (experimental culture). After cells had been grown at least for three generations, OD₆₀₀ was measured around every half doubling of cell growth. At each time point, OD₆₀₀ was measured by collecting 200 µl cell culture in a cuvette (Starna Cells) and using a spectrophotometer (Thermo Scientific). About 4–6 OD₆₀₀ data points within the range 0.04 to 0.3 were used for calculating growth rate. All of the growth rates measured in this study are summarized in Supplementary Table 2.

Measurement of expansion speeds

Expansion speeds were measured using soft-agar plates containing 0.25% agar and growth media resembling the liquid culture conditions described above. Attractants were additionally provided and a detailed preparation protocol is provided in Supplementary Text 1.4. Expansion speeds were measured either by manual tracking of expanding ring positions over time (manual observation, ring position is clearly visible by eye) or by using confocal microscopy. For the manual observation, 15 ml of freshly prepared and still warm soft-agar medium was transferred

into a Petri dish with a 10-cm diameter, resulting in a 2-mm-thick soft-agar layer. Agar was left to solidify for a minimum of 10 min at room temperature. For the confocal experiments, GFP-expressing plasmids were used as fluorescent markers, strains as indicated in the legends. Chloramphenicol (8 µg/ml) was additionally supplied into the soft-agar medium to maintain the plasmid. To prepare the soft-agar plates, 2.7 ml of the medium was transferred into a glass-bottom Petri dish (Ted Pella Inc.). The final thickness of the agar was approximately 1.2 mm. All plates were freshly prepared before the assay. To start the soft-agar assay (manual observation), 2 µl of cell culture from the (exponentially growing) experimental culture was transferred onto a pre-warmed soft-agar plate. The primary carbon source of the liquid culture was chosen such that it matched the growth conditions provided in the soft-agar plate. The plates were incubated at 37 °C. After the population covered a circular area of at least 2 cm in radius, the radius of the population (the front with chemotactic ring is clearly visible) was measured every 1–2 h for 4–6 time points. Expansion speeds were obtained as linear fits of the observed radii versus observation times (Fig. 1b). For convenience, the initial inoculation OD₆₀₀ used for the manual assay was varied depending on culture conditions such that ring movements could be captured during the day. The expansion speeds examined in this study are listed in Supplementary Tables 3, 9, 10. For expansion observation and density scans using microscopy, cells were always inoculated at OD₆₀₀ = 0.2 and observation was started immediately after the 2-µl cell culture was added to the agar. Expansion dynamics was analysed by looking at the emerging spatial intensity profiles. Details of confocal imaging, intensity analysis and calibration, and the determination of growth rates within soft-agar by confocal microscopy are provided in Supplementary Text 1.4.2–1.4.5. Custom-made code used is available via GitHub at https://github.com/jonascremer/chemotaxis_imageanalysisexpansiondynamics.

Measurements of swimming characteristics

To quantify the swimming behaviour of cells in liquid culture (no gradients), we grew cells as described above but with higher attractant concentrations to avoid the formation of temporal gradients (Supplementary Text 1.3.1). Polyvinylpyrrolidone was added to the medium to prevent cells from binding to surfaces and to protect flagella⁵⁵. Sample volumes of about 200 µl were taken at different time points over the course of cell growth to quantify swimming behaviour (see Supplementary Text 1.3.1 for details on timing). Immediately after collection, samples were diluted to a lower OD₆₀₀ of approximately 0.005 using filtered medium. The diluted sample was then used to load a rectangular capillary and cells within the capillary were observed by acquiring videos for 1 min, using a phase-contrast microscope. Using a custom-made Python script, cells were detected and cell trajectories were derived. Subsequently, cell trajectories were analysed to derive the swimming characteristics as previously reported⁵⁶. Full details on data acquisition and analysis are provided in Supplementary Text 1.3. The code used is available via GitHub at https://github.com/jonascremer/swimming_analysis.

Cell trajectory analysis in soft agar

To quantify diffusion behaviour (undirected run and tumbling) and drift (directed run and tumbling) of swimming cells within the agar, we used time-lapse confocal microscopy to enable the detection and tracking of individual fluorescently labelled cells (Extended Data Fig. 3). The measurement enables the spatiotemporal resolution of swimming behaviour within an expanding population. To optimize the tracking of single cells, the number of fluorescent detectable cells was adjusted by mixing fluorescent cells with non-fluorescent cells (carrying a non-fluorescence protein⁵⁷ to minimize physiological differences between strains). Detailed methods on image acquisition, cell tracking and the statistical analysis to derive diffusion coefficients and drift are given in Supplementary Text 1.5.

Measurements of aspartate uptake

To quantify aspartate uptake, cells were grown in minimal medium supplemented with different carbon sources and 800 μ M aspartate. Samples were collected at different optical densities during steady-state growth and the aspartate concentration was determined using a calorimetric aspartate kit (ab102512, Abcam). Additional details are provided in Supplementary Text 1.2.2. The aspartate consumption rates measured in this study are summarized in Supplementary Table 7. For each growth condition, measurements were repeated twice.

The growth–expansion model

Full details of the growth–expansion model introduced in Fig. 3 are provided in Supplementary Text 2. This includes the specific biological motivations for different terms used in the equations, including the terms that describe attractant consumption and the nutrient-dependent local growth rate (Monod type dependence^{58–60}). The drift term that describes the directed movement along sensed gradients features Weber's law^{7,61}, and response rescaling^{62–64} and parameters were taken from published receptor characterizations^{21,23,65,66}. Reflecting boundary conditions and initial conditions matching the experimental conditions were used; see Supplementary Text 2 for details and equations. Numerical solution of the partial differential equations used an implicit scheme using Python and the module FiPy⁶⁷. Integration over time was performed with time steps $dt = 0.25$ s, and a grid resolution with spacing $dx = 10$ μ m. Simulations were performed using custom-made Python code, which is available via GitHub at https://github.com/jonascremer/chemotaxis_simulation. Parameters used are provided in Supplementary File simulationparameters.txt; see also Supplementary Text 2.5 for additional information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Major experimental data that support this study are provided in this manuscript or available via figshare repositories: https://figshare.com/articles/Confocal_intensity_scans_expanding_bacteria/9639209 (confocal expansion data) and https://figshare.com/articles/Swimming_in_liquid_culture/9643001 (swimming observation data). Simulation data can be generated with the provided simulation code. Simulation parameters are provided in a separate file. The Supplementary Text provides additional details on strains, experimental methods and modelling.

Code availability

Custom-made code is available via GitHub for the analysis of swimming characteristics (https://github.com/jonascremer/swimming_analysis), the analysis of expanding populations using confocal microscopy (https://github.com/jonascremer/chemotaxis_imageanalysisexpansiondynamics), and the numerical simulations of the growth–expansion model (https://github.com/jonascremer/chemotaxis_simulation).

49. Lyons, E., Freeling, M., Kustu, S. & Inwood, W. Using genomic sequencing for classical genetics in *E. coli* K12. *PLoS ONE* **6**, e16717 (2011).
50. Soupene, E. et al. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J. Bacteriol.* **185**, 5611–5626 (2003).
51. Brown, S. D. & Jun, S. Complete genome sequence of *Escherichia coli* NCM3722. *Genome Announc.* **3**, e008795 (2015).
52. Basan, M. et al. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature* **528**, 99–104 (2015).

53. Parkinson, J. S. Complementation analysis and deletion mapping of *Escherichia coli* mutants defective in chemotaxis. *J. Bacteriol.* **135**, 45–53 (1978).
54. Cayley, S., Record, M. T., Jr & Lewis, B. A. Accumulation of 3-(N-morpholino) propanesulfonate by osmotically stressed *Escherichia coli* K-12. *J. Bacteriol.* **171**, 3597–3602 (1989).
55. Berg, H. C. & Turner, L. Chemotaxis of bacteria in glass capillary arrays. *Escherichia coli*, motility, microchannel plate, and light scattering. *Biophys. J.* **58**, 919–930 (1990).
56. Masson, J.-B., Voisinne, G., Wong-Ng, J., Celani, A. & Vergassola, M. Noninvasive inference of the molecular chemotactic response using bacterial trajectories. *Proc. Natl Acad. Sci. USA* **109**, 1802–1807 (2012).
57. Liu, W., Cremer, J., Li, D., Hwa, T. & Liu, C. An evolutionarily stable strategy to colonize spatially extended habitats. *Nature* <https://doi.org/10.1038/s41586-019-1734-x> (2019).
58. Shehata, T. E. & Marr, A. G. Effect of nutrient concentration on the growth of *Escherichia coli*. *J. Bacteriol.* **107**, 210–216 (1971).
59. Schellenberg, G. D. & Furlong, C. E. Resolution of the multiplicity of the glutamate and aspartate transport systems of *Escherichia coli*. *J. Biol. Chem.* **252**, 9055–9064 (1977).
60. Cremer, J. et al. Effect of flow and peristaltic mixing on bacterial growth in a gut-like channel. *Proc. Natl Acad. Sci. USA* **113**, 11414–11419 (2016).
61. Shimizu, T. S., Tu, Y. & Berg, H. C. A modular gradient-sensing network for chemotaxis in *Escherichia coli* revealed by responses to time-varying stimuli. *Mol. Syst. Biol.* **6**, 382 (2010).
62. Shoval, O. et al. Fold-change detection and scalar symmetry of sensory input fields. *Proc. Natl Acad. Sci. USA* **107**, 15995–16000 (2010).
63. Lazova, M. D., Ahmed, T., Bellomo, D., Stocker, R. & Shimizu, T. S. Response rescaling in bacterial chemotaxis. *Proc. Natl Acad. Sci. USA* **108**, 13870–13875 (2011).
64. Celani, A., Shimizu, T. S. & Vergassola, M. Molecular and functional aspects of bacterial chemotaxis. *J. Stat. Phys.* **144**, 219–240 (2011).
65. Vaknin, A. & Berg, H. C. Physical responses of bacterial chemoreceptors. *J. Mol. Biol.* **366**, 1416–1423 (2007).
66. Neumann, S., Grosse, K. & Sourjik, V. Chemotactic signaling via carbohydrate phosphotransferase systems in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **109**, 12159–12164 (2012).
67. Guyer, J. E., Wheeler, D. & Warren, J. A. FiPy: partial differential equations with Python. *Comput. Sci. Eng.* **11**, 6–15 (2009).
68. Dufour, Y. S., Gillet, S., Frankel, N. W., Weibel, D. B. & Emonet, T. Direct correlation between motile behavior and protein abundance in single cells. *PLOS Comput. Biol.* **12**, e1005041 (2016).
69. Keestra, J. M. et al. Phenotypic diversity and temporal variability in a bacterial signaling network revealed by single-cell FRET. *eLife* **6**, e27455 (2017).
70. Frankel, N. W. et al. Adaptability of non-genetic diversity in bacterial chemotaxis. *eLife* **3**, e03526 (2014).
71. Müller, M. J. I., Neugeboren, B. I., Nelson, D. R. & Murray, A. W. Genetic drift opposes mutualism during spatial population expansion. *Proc. Natl Acad. Sci. USA* **111**, 1037–1042 (2014).
72. Möbius, W., Murray, A. W. & Nelson, D. R. How obstacles perturb population fronts and alter their genetic structure. *PLOS Comput. Biol.* **11**, e1004615 (2015).
73. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nat. Commun.* **7**, 12760 (2016).
74. Weinstein, B. T., Lavrentovich, M. O., Möbius, W., Murray, A. W. & Nelson, D. R. Genetic drift and selection in many-allele range expansions. *PLOS Comput. Biol.* **13**, e1005866 (2017).
75. Mesibov, R., Ordal, G. W. & Adler, J. The range of attractant concentrations for bacterial chemotaxis and the threshold and size of response over this range. Weber law and related phenomena. *J. Gen. Physiol.* **62**, 203–223 (1973).
76. Fraebel, D. T. et al. Environment determines evolutionary trajectory in a constrained phenotypic space. *eLife* **6**, e24669 (2017).
77. Blattner, F. R. et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).

Acknowledgements We thank C. Liu and X. Fu for initiating this study, and H. Berg, P. Cluzel, K. Fahrner, J. S. Parkinson, T. Pilizota, T. Shimizu, V. Sourjik and Y. Tu for discussions. This work is supported by the NIH (R01GM95903) through T. Hwa and the NSF Program PoLS (grant 1411313) through M.V. T. Honda acknowledges a JASSO long-term graduate fellowship award.

Author contributions J.C., T. Honda, M.V. and T. Hwa designed this study. Experiments were performed by T. Honda and J.C., with contributions by J.W.-N. and Y.T. in characterizing swimming. J.C. and T. Hwa developed the model, and J.C. and Y.T. performed the numerical simulations. All authors contributed to the analysis of experimental and simulation data. J.C., T. Honda, Y.T., M.V. and T. Hwa participated in the writing of the paper and the Supplementary Information.

Competing interests The authors declare no competing interests.

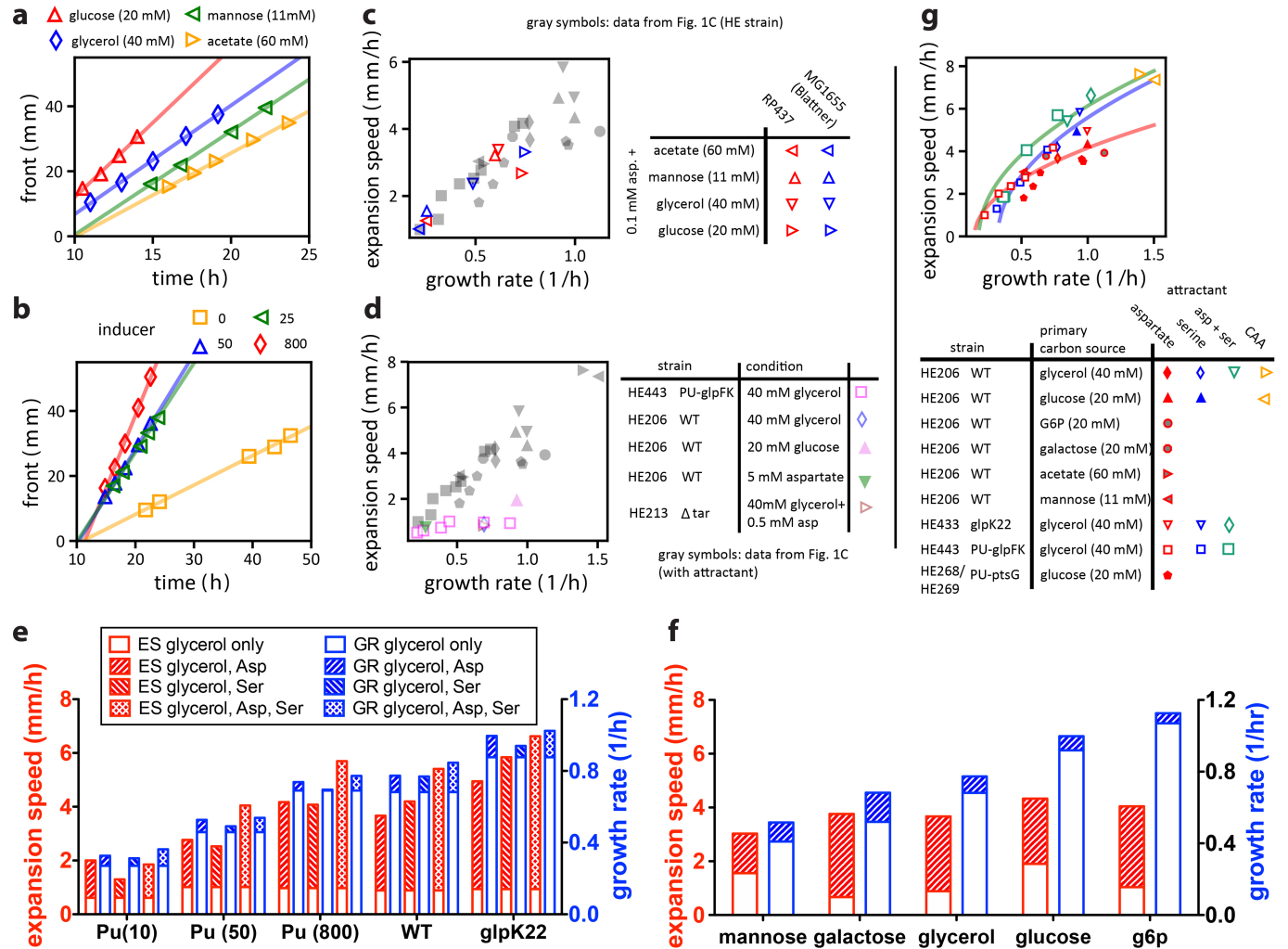
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1733-y>.

Correspondence and requests for materials should be addressed to T.H.

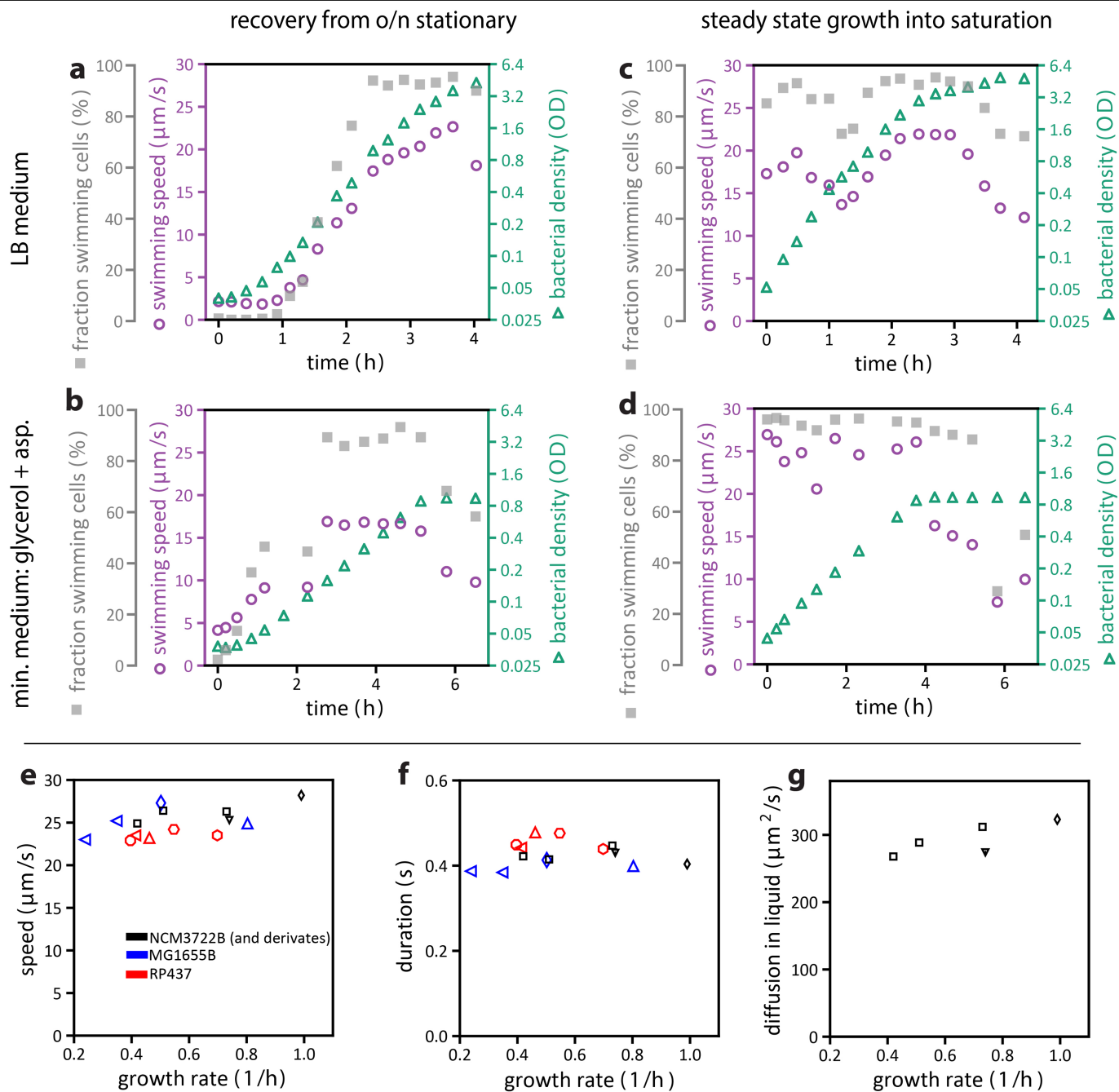
Reprints and permissions information is available at <http://www.nature.com/reprints>.

Article



Extended Data Fig. 1 | Expansion speed measurements. **a**, Temporal evolution of front position for a population of *E. coli* HE206 cells (wild type) grown on a soft-agar plate with saturating amounts of different carbon sources (see legend) and 100 μ M aspartate as attractant. Lines show linear fits. Experiments were repeated at least twice with similar results. **b**, Temporal evolution of front position for HE443 cells grown on 40 mM glycerol and 100 μ M aspartate with different amounts of the inducer 3MBA (see legend) that titrate glycerol uptake²⁶, resulting in different growth rates (Supplementary Table 2). The experiments were repeated at least twice with similar results. **c**, Expansion speed and its dependence on growth rate for the commonly used *E. coli* K-12 strain MG1655⁷⁷ (red symbols) and the K-12 variant RP437 (blue symbols) frequently used in motility studies (see Supplementary Text 1.1). Growth conditions were changed by varying the carbon source (from lower to higher growth rates: acetate, mannose, glycerol, glucose; see legend). Aspartate (100 μ M) was added as the attractant. For the experiments with RP437, four amino acids (methionine, leucine, threonine, histidine) were provided at 1 mM each in the medium to sustain cell growth. Data from Fig. 1c are shown in grey for comparison. Data points represent the mean of two biological replicates, except for growth rates in acetate and mannose, which were from a single experiment. **d**, Expansion speeds plotted against the batch culture growth rate for populations grown in glycerol, glucose or aspartate as the only carbon source, without supplement of additional attractant (purple symbols). Growth

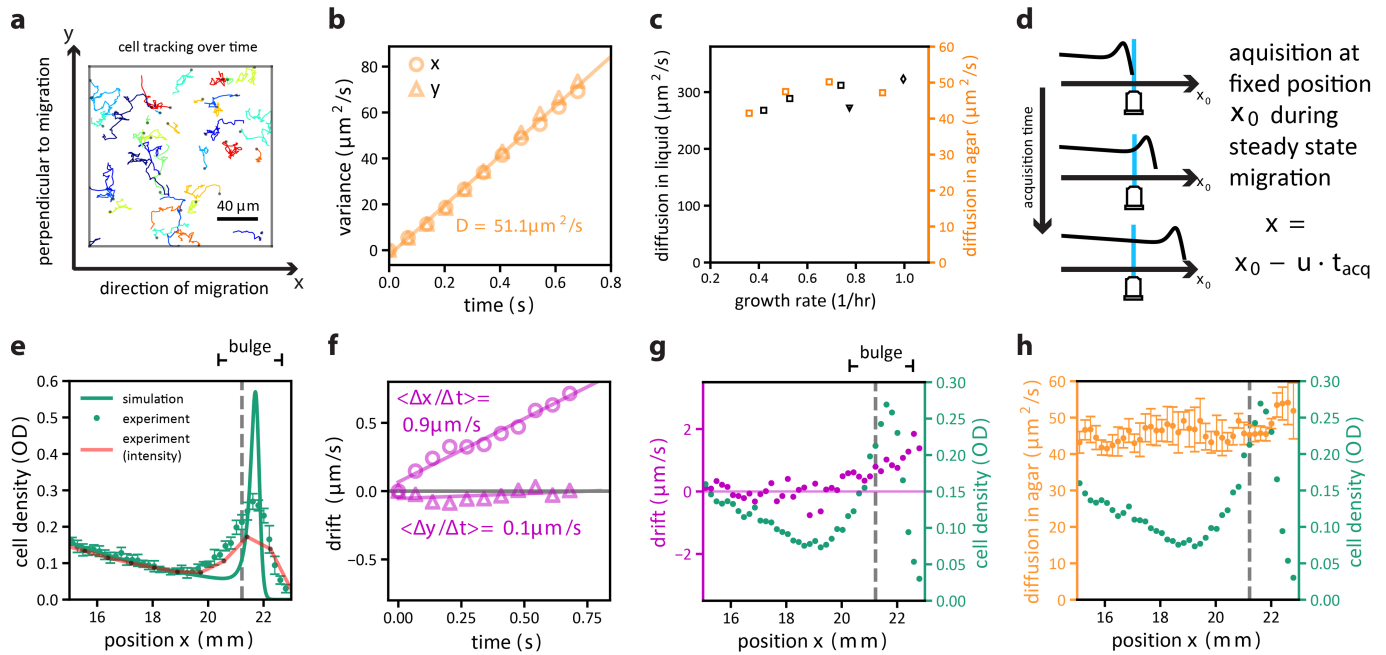
on serine is very slow ($<0.1 \text{ h}^{-1}$) and is not shown on the plot. Expansion speeds were much slower in these media without chemoattractant supplement, even though glucose and aspartate are both attractants themselves^{21,45}. The same was observed for a Δtar knockout strain when both glycerol and aspartate were present (open triangle). See Supplementary Tables 2, 3 for data values and sample sizes. Data from Fig. 1c are shown in grey for comparison. **e**, **f**, The difference between migration with and without additional attractant is further illustrated for growth on glycerol when growth rates are titrated (**e**), and for expansion when other carbon sources are provided (**f**). Hashed bars highlight additional increase in expansion speed (red) and growth rate (blue) when attractant is provided. In each case, supplementing low amounts of attractant(s) greatly increases expansion speed without affecting growth rate much. The graphs were created based on mean values listed in Supplementary Tables 2, 3. **g**, Expansion speed and its dependence on growth rate when two attractants are present (20 mM glycerol + 100 μ M aspartate + 100 μ M serine, green symbols) or for complex medium (CAA + carbon source, orange symbols). Data for single attractants (from Fig. 1c) are shown for comparison (20 mM glycerol + 100 μ M aspartate, red; 20 mM glycerol + 100 μ M serine, blue). Lines indicate fits to square-root dependencies as anticipated from a simple scaling analysis (Extended Data Fig. 9d). Data points in asp + ser (green points) represent means of two biological replicates ($n = 2$), except for growth rates of HE433 and HE443, which were from a single experiment ($n = 1$).



Extended Data Fig. 2 | See next page for caption.

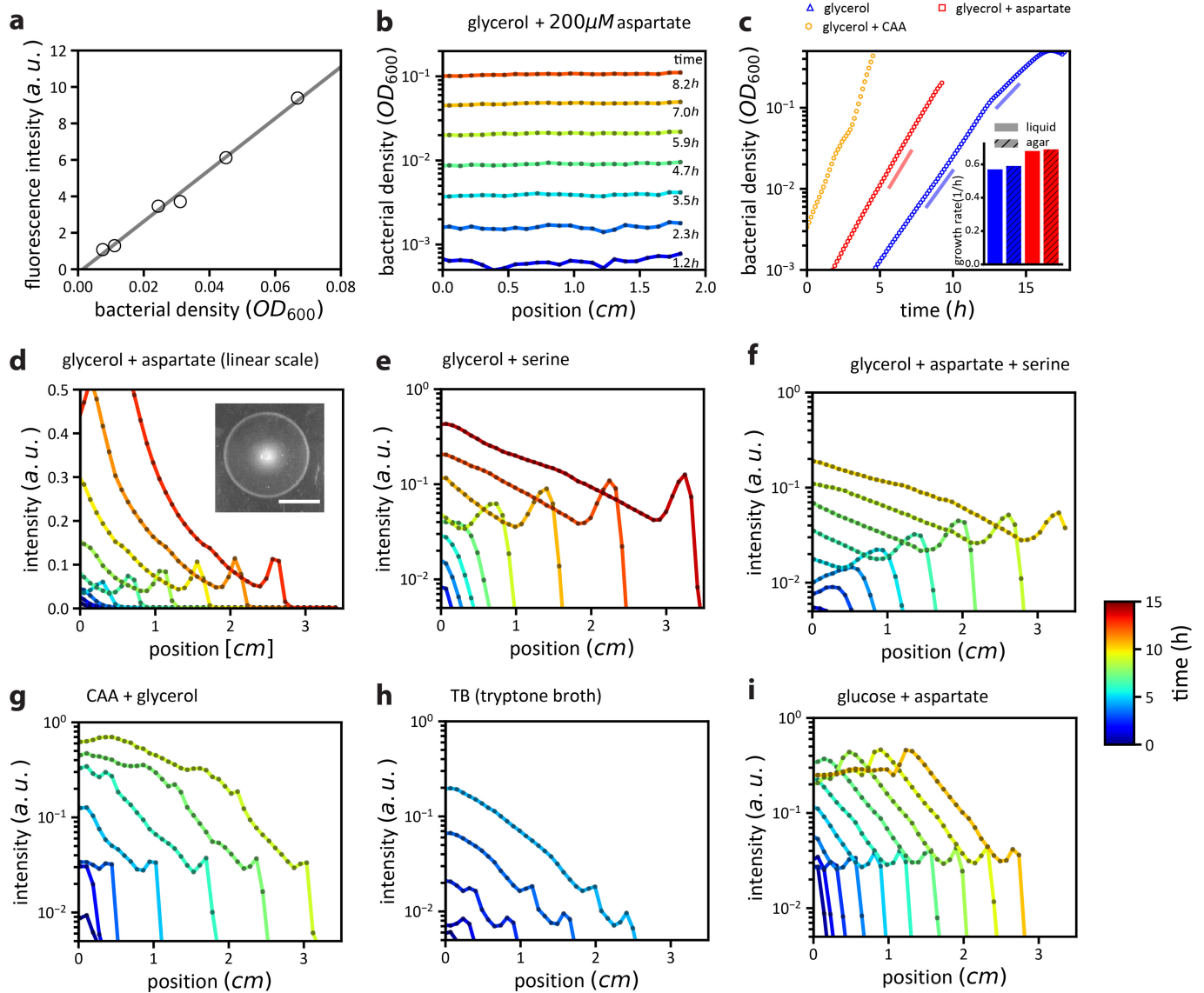
Extended Data Fig. 2 | Swimming characteristics in liquid media (well-mixed conditions, no gradients). a–d, Average swimming speed (purple circles) and the fraction of motile cells (grey squares) were characterized for cells taken from batch cultures along a growth curve at different optical densities (OD_{600} ; green triangles). For each condition, data points were collected from a single experiment. **a,** Culture was grown in LB medium, starting with an overnight LB culture that was sitting in saturation for 18 h before dilution into fresh LB medium at time zero. This experiment was essentially a repeat of previous work^{14,15}; similar results were obtained, with motility increasing as growth progressed. Our data in the following panels suggest that most of the increase in swimming speed resulted from the increased fraction of motile cells in the first 2 h. **b,** Culture was grown in minimal medium with 10 mM glycerol and 1.7 mM aspartate, starting with an overnight pre-culture (same medium) that was in saturation for about 18 h before inoculation into fresh medium (time zero). As observed for LB (**a**), it took several hours for both the motile fraction and swimming speed to recover. **c,** Culture was grown in LB medium continuously for 10 generations, with bacterial density always kept below $OD_{600} = 0.5$ before dilution to fresh LB at time zero. Both the motile fraction and the swimming speeds are high in the exponential growth phase (0–2 h) except for a dip at an OD_{600} of approximately 0.5. Swimming speed and motile fraction decreased after the stationary phase was reached. **d,** Culture was grown in the

same minimal medium (glycerol + aspartate) for about 20 generations, with bacterial density maintained below $OD_{600} = 0.6$ before measurement. As with LB (**c**), swimming speed and motile fraction remained high in the exponential growth phase (0–4 h) before sharply decreasing after entering the stationary phase. The strong variation of the fraction of motile cells observed here is in line with previous observations on cell-to-cell variation in swimming behaviour^{68,69} and can strongly affect the dynamics of migrating populations^{24,70}. **e, f,** Swimming behaviour observed in steady-state growth (as in **d** for the first ~3 h) for different (relatively fast) growth conditions and different *E. coli* strains (Supplementary Table 5). Swimming speeds (v) and durations between tumbling events (τ) obtained from trajectory analysis are shown in **e** and **f**, respectively. Black symbols show results for the NCM3722B-derived strains (HE206, HE433, HE443; growth at 37 °C) mainly used in this study. A similar weak dependence of quantities on growth was observed for MG1655B (red symbols, growth at 37 °C) and RP437 (blue symbols, growth at 30 °C). **g,** Estimated effective diffusion coefficient, $D = v^2\tau$, for the different growth conditions in NCM3722B-derived strains. **e–g,** Data points show the means of two biological replicates for HE433 and HE443 and the results of single experiments for HE206, MG1655B and RP437. See Supplementary Text 1.2, 1.3 for methods, Supplementary Table 5 for data values and conditions, and Supplementary Text 1.1 for strain details.



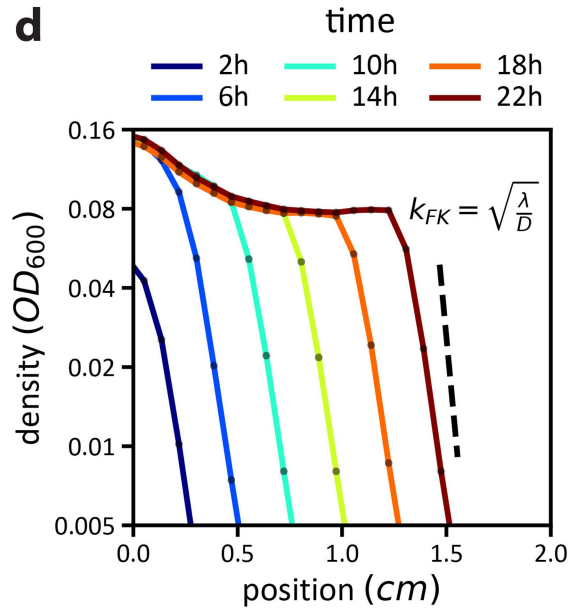
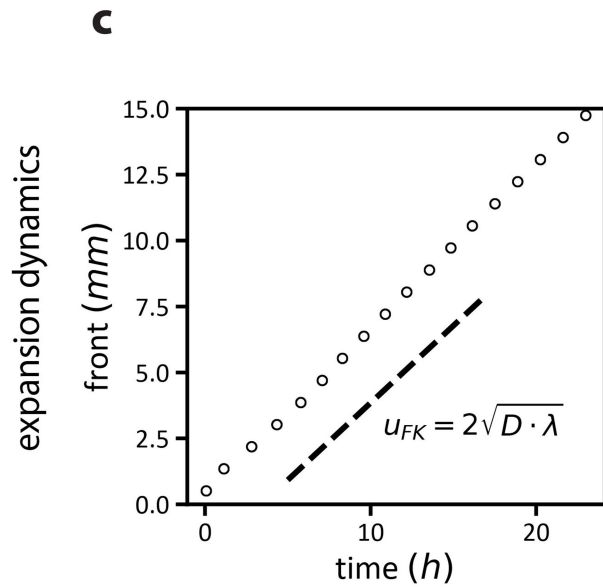
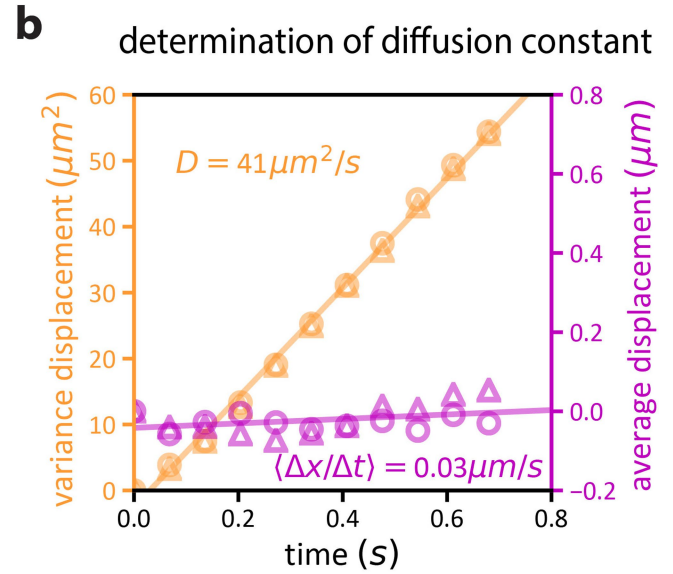
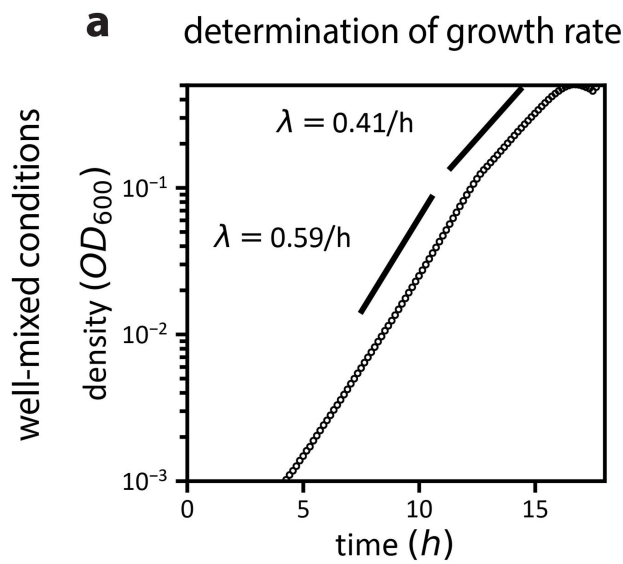
Extended Data Fig. 3 | Single-cell motility analysis in agar by confocal microscopy. Thirty-second videos allowing us to track the movement of single cells were acquired (see Supplementary Video 7 for an example). **a**, Example of trajectories derived from cell tracking analysis. Each colour indicates the trajectory of one cell over a span of on average 75 frames (5.1 s). **b**, Diffusive behaviour was obtained by a linear fit of displacement variance over time ($\text{var}(\Delta x) = 2D\Delta t$). This analysis was performed for strain HE274 (wild type) growing in 40 mM glycerol and 100 μM aspartate (reference condition; Supplementary Text 1.5). Data shown here are for measurements in front of the expanding population (ahead of the density peak; however, the diffusion coefficient obtained at different locations does not exhibit much positional dependency, see below). Repeat of experiment showed similar results. **c**, Similar effective diffusion coefficients for swimming in soft-agar were obtained for other growth conditions (orange symbols; Supplementary Table 6) following the same trend as predicted from liquid culture measurements (black symbols, same as in Extended Data Fig. 2g). The diffusion measurements in soft agar were repeated twice with similar results. The data points represent means of two biological replicates. See Supplementary Table 6 for data values and conditions. **d**, To resolve cellular swimming behaviour of the expanding population at different spatial positions in the agar plate, videos allowing us to track single cells were acquired sequentially at a fixed position (of the agar plate) over time, for different acquisition times t_{acq} over which videos were taken (up to several hours for each position). Image direction x was aligned with direction of migration. In this setup, the migrating population (with speed u) passes the point of acquisition at a determined time, allowing us to determine the local drift speeds and diffusion coefficients relative to the front position: $x = x_0 - ut_{\text{acq}}$ (Supplementary Text 1.5). **e**, Density

obtained by cell counting (green line) compared to population density obtained using the approach in Extended Data Fig. 4 (fluorescence scans, red line). The spatial resolution of the latter is much coarser, each measurement point being a black dot on the red line. For comparison, the simulation result (GM model, Fig. 3) is shown in green and moderately deviates from the measured profile. **f**, Analysis of average displacement along x (direction of migration) and y (direction perpendicular to migration) over time for an acquisition time t_{acq} corresponding to a position at the front bulge ($x = 21.3$ cm, indicated by the dashed lines in **e**, **g**, **h**). The average displacement (purple symbols) increased linearly in time along the direction of migration but was negligible perpendicular to the direction of migration (fitted purple lines show drift speed in each direction, $\langle \Delta x / \Delta t \rangle$ and $\langle \Delta y / \Delta t \rangle$). **g**, Position dependence of the drift (in the direction of expansion) was determined at different t_{acq} , corresponding to different positions of the expanding population. For ease of reference, cellular densities (**e**) are shown again as green symbols. Up to the resolution of the data, the drift velocity vanished to the left of the density trough ($x < 19$ mm). **h**, Position dependence of the diffusion coefficient. Using the approach from **b** to determine the diffusion coefficient at different t_{acq} , we obtained the results shown as orange symbols. A moderate (~20%) increase in D is observed at the very front of the population. This spatial dependence may be due to the accumulation of faster swimming cells at the front⁹. All data in **e–h** are from a single expansion experiment done under reference conditions (40 mM glycerol + 100 μM aspartate; 2:1 mixture of fluorescent variant HE274 and non-fluorescent variant HE339). Similar results were obtained for one biological replicate. Error bars in **e**, **h** denote s.d. and were calculated from repeated observations at three different times during the same expansion process.



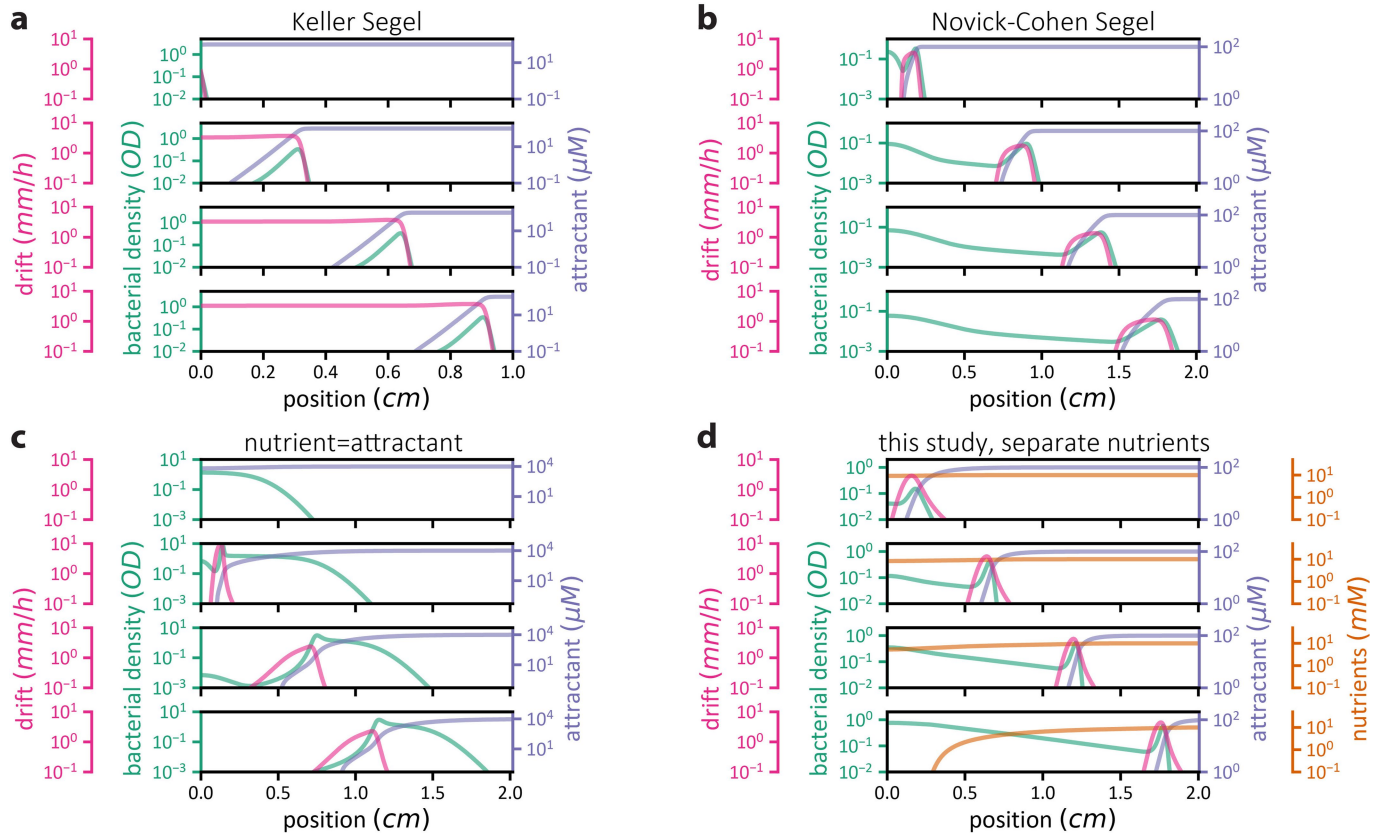
Extended Data Fig. 4 | Population-level observations of growth and expansion by confocal microscopy. Densities of bacteria growing in soft agar were determined at various times using confocal microscopy and fluorescently labelled cells; see Supplementary Text 1.4. **a**, Calibration of fluorescence intensity. Known numbers of cells were transferred from a batch culture to a fresh but cold soft-agar plate. After agar solidification (<10 min), intensity was measured. Fitted line gives relation between observed intensity (fluorescence integrated along the agar thickness) and cell density measured in batch culture (OD_{600}). **b**, Example of experiment to obtain growth rates in agar. Data are for strain HE274 (wild-type) grown in 40 mM glycerol and 200 μ M aspartate. A small number of cells was mixed uniformly into fresh soft-agar plates. After agar solidification (<10 min), fluorescence intensity was observed over time. **c**, Derived growth curves in soft agar based on experiments as in **b**. Typically, there is a fast growth regime followed by a slower regime related to oxygen consumption and limitation for $OD > 0.1$: with oxygen running out, cells accumulate towards the agar surface and growth becomes slower. In this work,

the population was always kept in the first aerobic regime. Growth rates in the first regime (coloured lines, Supplementary Table 2) were obtained by an exponential fit of the data and are comparable to those obtained in batch culture (inset). In **b**, **c** the experiment was conducted once. **d**, Photograph and spatiotemporal density profiles (linear intensity scale) for population expansion under reference conditions (glycerol + 100 μ M aspartate; same data as in Figs. 1a, 2b). Scale bar, 2 cm. The confocal observations under reference conditions were repeated twice with similar results. **e–i**, Spatiotemporal density profiles (logarithmic density scale) for population expansion under different conditions, similar to those observed for the reference conditions (glycerol + 100 μ M aspartate; **d**). Conditions are glycerol + 100 μ M serine (**e**), glycerol + 100 μ M aspartate + 100 μ M serine (**f**), glycerol + 0.05% CAA (**g**), 1% tryptone broth (**h**), and glucose + 100 μ M aspartate (**i**), all with strain HE274 (wild type). Colour scale bar applies to all panels. In **e–i**, the experiments were conducted once; expansion speeds were highly comparable to those measured manually.



Extended Data Fig. 5 | Population expansion without attractant is quantitatively captured by Fisher–Kolmogorov dynamics. The Fisher–Kolmogorov dynamics is a canonical model to describe the dynamics of expanding populations^{19,20}. It has been successfully used to investigate the expansion and evolution of non-moving bacteria at the front of dense bacterial colonies^{46,71–74}. Here, we probe the Fisher–Kolmogorov dynamics and its validity to describe swimming bacteria. The Fisher–Kolmogorov dynamics is driven by population growth and undirected random motion (diffusion)^{32,33}. To compare the predictions of Fisher–Kolmogorov dynamics to the expansion of a bacterial population in the absence of a chemoattractant, we thus independently quantified growth rates and cellular diffusion for cells homogeneously distributed in soft agar (**a**, **b**). We then compared the observed migration speed and the density profile of the migrating population (for growth on glycerol as the sole carbon source, as in Fig. 2d, top) with the Fisher–Kolmogorov predictions (**c**, **d**). **a**, Quantification of growth by measuring the temporal density increase of a homogeneously distributed population in agar (Extended Data Fig. 4a–c, Supplementary Text 1.4). Spatially averaged density increased exponentially with growth rate $\lambda = 0.59 h^{-1}$ for densities $< 0.1 OD_{600}$. For higher densities, the growth rate decreased but this regime is not important for the propagation of the front where density is low. **b**, Diffusion and drift of cells homogeneously distributed in soft agar. Analysis of recorded cell movement confirms the variance of position displacement to increase linearly in time

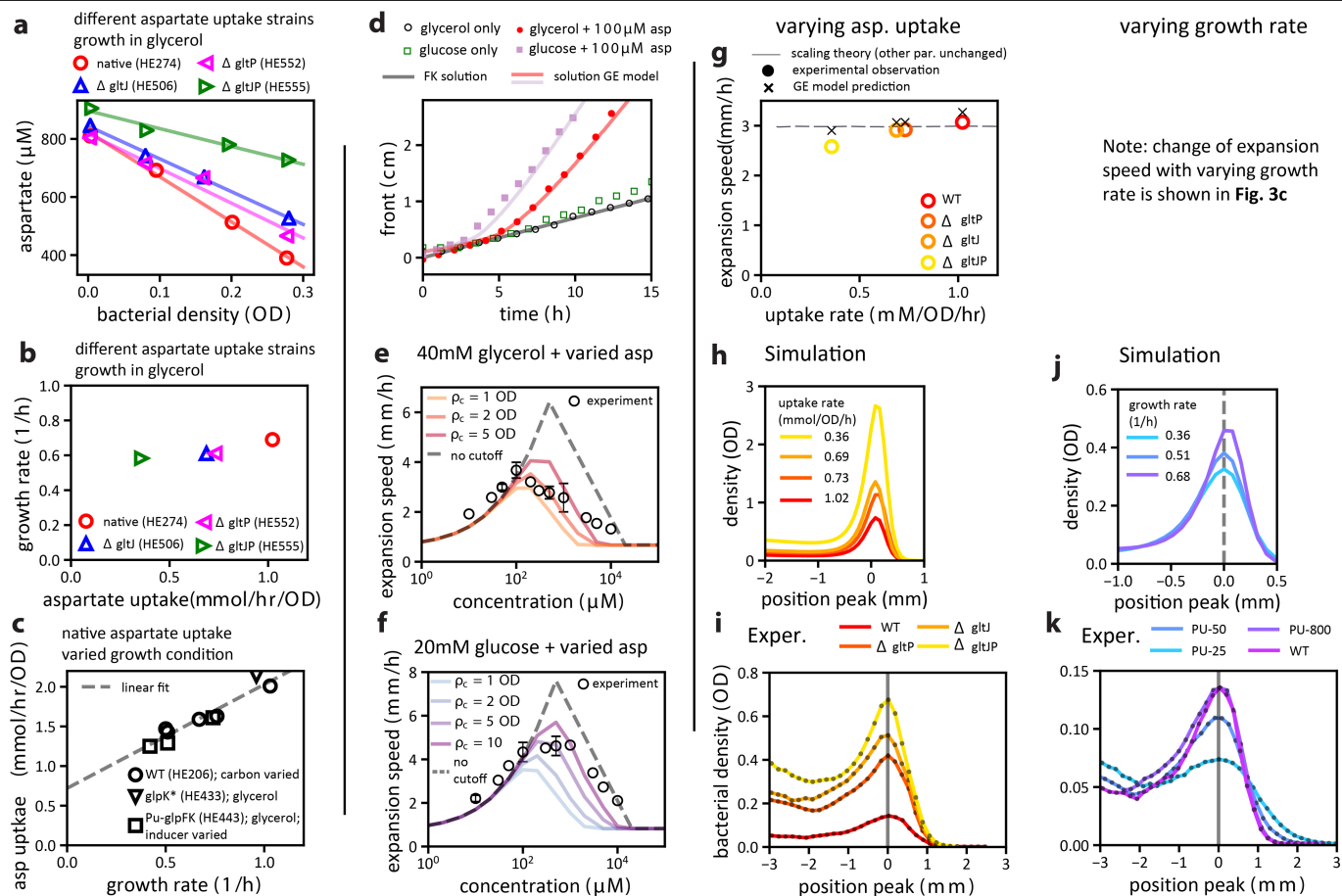
(orange symbols) with diffusion constant $D = 41.5 \mu m^2 s^{-1}$ (linear fit of $var(x) = 2D\Delta t$). In comparison, the average displacement of cells (purple symbols) and the calculated drift ($\langle \Delta x / \Delta t \rangle$, purple line) are small, indicating the absence of directed chemotactic movement. Data show average over three independent repeats (Extended Data Fig. 3, Supplementary Text 1.5). **c**, **d**, Front and spatiotemporal dynamics of an expanding population. **c**, Comparison of predicted expansion speed with the observed propagation of the population front. Position of the front $R(t)$ was determined from the observed cellular densities (threshold $OD_{600} < 0.005$); it increased linearly in time, that is, $R(t) = u_{obs} t$ with a speed $u_{obs} = 0.62 mm h^{-1}$. Dashed line denotes predicted expansion speed calculated as $u_{FK} = 2\sqrt{\lambda \times D} = 0.59 mm h^{-1}$. **d**, Density profile of the population front. Observed density profile can be fitted to an exponential dependence $\rho(r, t) \sim e^{-k_{obs}(r-R(t))}$ with $k_{obs} \approx 1.2 mm^{-1}$. Dashed line indicates the slope of the exponential density profile predicted by the Fisher–Kolmogorov equation: $k_{FK} = \sqrt{\lambda / D} = 1.99 mm^{-1}$. The discrepancy is likely to result from the low spatial resolution of the very sharp density drop; the exponential dependence of the experimental profile is defined by just three points. All experiments were conducted once with strain HE274 (wild type), using glycerol as the carbon source (no additional attractant, glycerol cannot be sensed). Growth and cell-tracking experiments were performed with uniform cell mixture in saturating glycerol conditions (40 mM). Expansion experiments were performed with 1 mM glycerol.



Extended Data Fig. 6 | Different models of chemotaxis-driven migration.

To illustrate the difference among various models of chemotactic expansion, we show here simulation results of four different models. **a**, The classical model proposed by Keller and Segel³⁴ creates a self-generated attractant gradient owing to attractant consumption by the migrating population. It neglects cell growth (that is, $\lambda = 0$ in equation (3) in Fig. 3a), resulting in conservation of the total number of bacteria. It also assumes that the attractant gradient could be detected with infinite precision, such that log-sensing (Weber's law⁷⁵) can be implemented by cells down to arbitrary low attractant concentrations, (that is, equation (4) with $a_- = 0$). The latter biologically unrealistic assumption introduces a singularity that pushes all bacteria forward at a steady migration speed, which is determined by the number of cells in the population, the conserved quantity. **b**, The model introduced by Novick-Cohen and Segel³⁶ fixed the singularity in the Keller-Segel model by imposing a minimal concentration for the sensing of attractant gradient (that is, equation (4) with $a_- > 0$). Owing to the lack of cell growth, the total number of bacteria is still conserved. In this model, the density of the front bulge decays over time because once bacteria diffuse out of the front, they lose the chemotactic

gradient and cannot catch up with the front. The reduction in front density reduces the migration speed, which decays steadily towards zero. **c**, Model including cell growth that depends on attractant concentration (nutrient = attractant). Owing to growth, population size increases over time. However, as the attractant (nutrient) is mostly consumed at the front, there is not much growth behind the front and the trailing region behind the front is mostly flat. This scenario has been realized and analysed experimentally¹⁸; see Extended Data Fig. 8 for model details and discussion. **d**, The GE formulated in this study (Fig. 3a), including the chemotactic effect of an attractant, together with cell growth supplied by a major nutrient source. Front propagation of cells by chemotaxis is coupled to steady growth in the trailing region (see main text and Extended Data Fig. 9). Parameter values for all models are provided in Supplementary Table 8. For simplicity, simulations shown here were solved in one dimension (non-radial). Green lines denote bacteria density, blue lines denote attractant (or sole nutrient) concentration, brown lines denote concentration of nutrients (in addition to the attractant), purple lines show local drift (equation (4)).

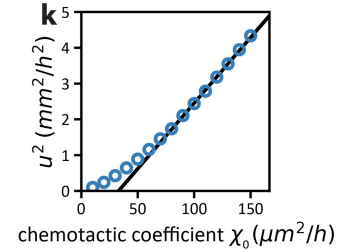
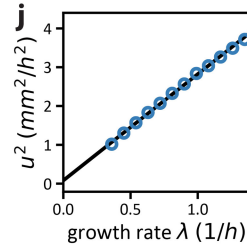
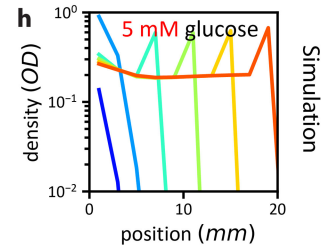
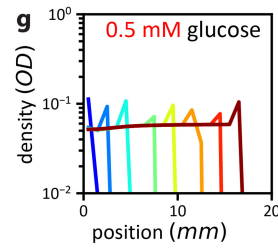
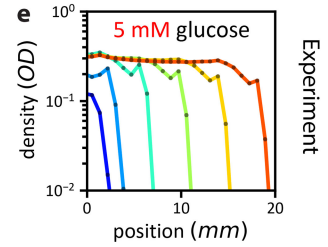
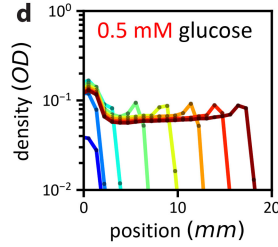
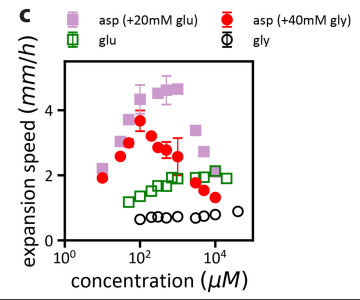
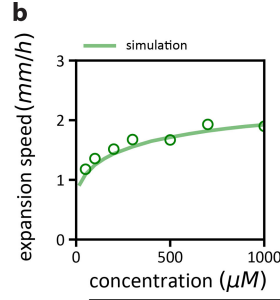
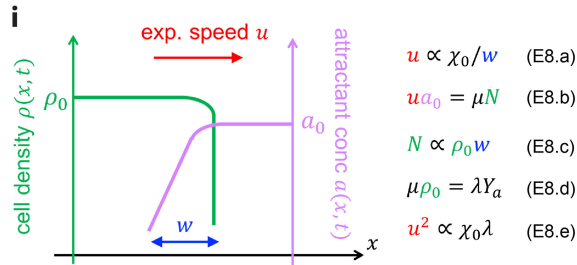
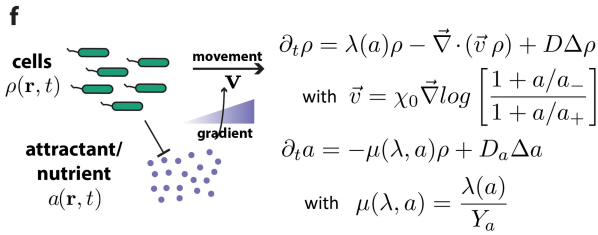
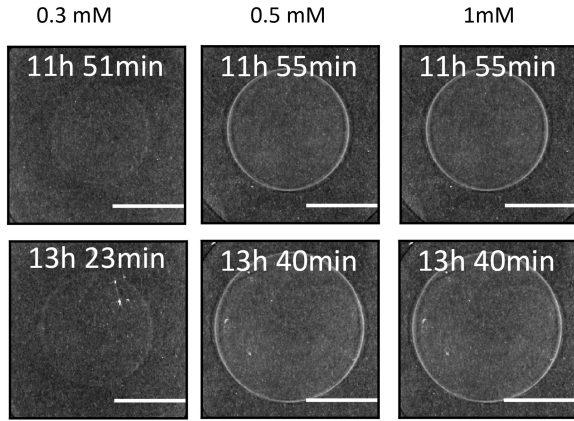


Extended Data Fig. 7 | Aspartate uptake and further analysis of expansion dynamics. **a–c**, Characterization of aspartate uptake for different growth conditions and strains. **a**, Aspartate uptake was determined using a colorimetric method to quantify remaining aspartate concentrations during growth (see Supplementary Text 1.2.3). In brief, change of aspartate concentration in the medium was measured during exponential growth at different cell densities (OD). Data are for growth with glycerol as the major carbon source (40 mM) and 0.8 mM initial aspartate concentration. Measurements for native aspartate uptake (wild type) and for different aspartate-uptake mutants (Δ gltJ and Δ gltP as well as the double mutant Δ gltJP). Lines show a linear fit. Uptake rate was determined by multiplying the observed slope by the growth rate. **b**, The strains shown in **a** with different aspartate uptake rates exhibit similar growth rates. For each strain, the data points were collected from a single experiment. **c**, Dependence of aspartate uptake on growth rate for strains with native aspartate uptake. Growth is varied using wild-type cells (HE206, circles) grown on different sugar sources (acetate, mannose, glycerol or glucose), or by using the *glpK** mutant (HE433, triangle) or the glycerol titration mutant (HE443, squares), in glycerol with different levels of the inducer 3MBA (25, 50 or 800 μ M). Aspartate (0.8 mM) was provided in each case for the measurement of aspartate uptake (see **a**). Line shows linear fit with parameters specified in Supplementary Text 1.2.3. **a, b**, Data obtained for strains carrying fluorescence plasmids. **c**, Data obtained for non-fluorescent strains (two biological replicates, means shown). Data, strain information and medium conditions including concentrations of carbon sources are provided in Supplementary Table 7. **d**, Expansion dynamics with glucose as the primary carbon source. The dynamics of the front, shown to be described well by the GE model in Fig. 3b in glycerol with aspartate, is examined with the primary carbon source being glucose (20 mM). In the presence of 100 μ M aspartate, the observed front propagation dynamics (purple squares) is correctly captured by the GE model again (purple line), by merely replacing the growth rate by that in glucose ($\lambda = 1.0$ h⁻¹) with no additional adjustment of the chemotactic coefficient χ_0 . For reference, expansion is also shown for the condition in which no additional chemoattractant was provided (0 μ M aspartate, open green squares), and the corresponding data for the condition in which glycerol was the primary carbon source (open black circles and corresponding lines from

Fig. 3b). **e, f**, Dependence of expansion speed on attractant concentration with glycerol or glucose being the major nutrient. The increase in expansion speed at low attractant concentrations followed by decrease at higher concentrations, as previously observed⁴³, is qualitatively captured by the GE model (dashed grey lines) in both cases. A better quantitative agreement between model and data is obtained when the linear growth term in the GE model (equation (1) in Fig. 3a) is changed to the logistic form $\lambda\rho(1 - \rho/\rho_c)$. Here ρ_c is the carrying capacity, introduced to capture saturation of cell density in the front bulge (Supplementary Text 2.3). Predictions by the model are shown for different carrying capacities as coloured lines. In line with the strict requirement for oxygen when growing on glycerol and the observation that cells at high density accumulate at the agar surface when expanding with glycerol as major nutrient source (data not shown), the carrying capacity needed to resemble the observations is much lower for glycerol (**e**) than for glucose (**f**). Data points represent means of biological replicates ($n = 2$ or more) with error bars (s.d.) shown for $n \geq 3$; see Supplementary Table 9 for data and sample sizes. **g–i**, Effect of varying aspartate uptake rate on expansion speed. The GE model predicts the expansion speed to be independent of the attractant uptake rate if all other parameters are kept fixed (**g**, dashed black line; Supplementary Text 2.2), with differences in attractant uptake compensated by changes in bacterial density at the front (**h**), such that the total rate of attractant depletion remains constant. This prediction was tested by characterizing the expansion dynamics of the aspartate-uptake mutants (strains HE506, HE552, HE555; Supplementary Text 1.1, Supplementary Tables 3, 7), which exhibited up to threefold difference in aspartate uptake (**a**), but only ~20% change in expansion speed (**g**). The small changes are readily accounted for by incorporating the small growth rate differences between these strains (**b**) into the GE model while keeping all other parameters fixed (**g**, green crosses). In addition, the aspartate-uptake mutants exhibited increasing peak densities at the front as predicted by compensation for reduced uptake (compare **i, h**). **j–k**, Predicted and observed changes in density profiles when varying the growth rate by titrating glycerol uptake in strain HE486 using 25 μ M, 50 μ M or 800 μ M of the inducer 3MBA. For **g, i, k**, data were obtained from a single experiment for each strain and condition.

Article

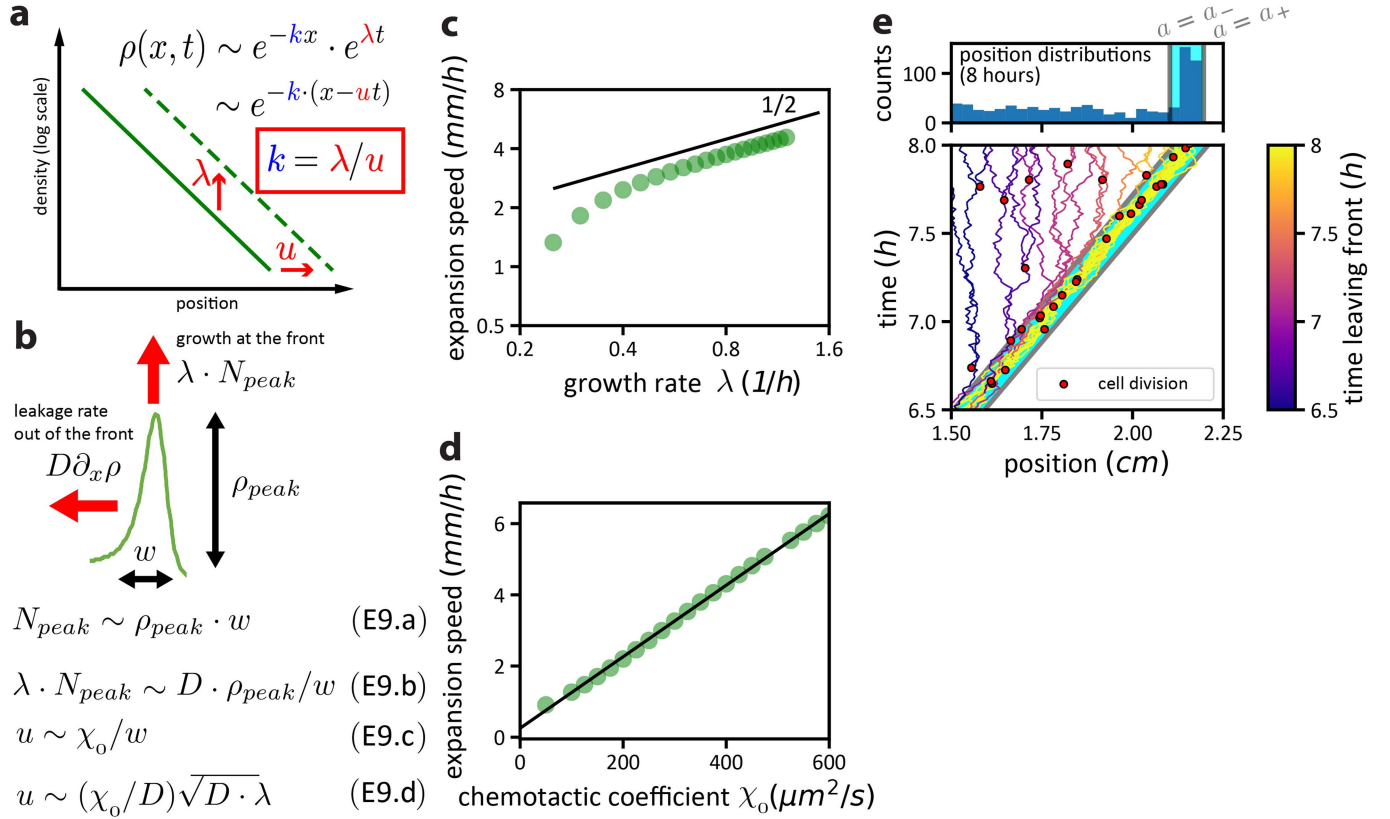
a expansion in glucose only



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Expansion dynamics with the attractant being the sole nutrient source. This is one of the scenarios of chemotaxis investigated previously^{18,76}. Here, for comparison with the dynamics presented in the main text, we show the expansion dynamics of populations grown with glucose (a chemoattractant²¹) as the sole carbon source. **a**, For wild-type cells (HE206) spotted on 0.25% agar plate with glucose as the sole carbon source, photographs show the existence of an outer ring at the front of the expanding population for a range of glucose concentrations. Scale bars, 2 cm. The experiments were repeated once with similar results. **b**, Dependence of expansion speed on glucose concentration. Intuitively, reducing the glucose concentration would be expected to increase the expansion speed, as it would take less time for the population to consume the attractant. However, the circles show that reducing the glucose concentration reduced population expansion speed. Data show means of two biological replicates. **c**, Direct comparison of concentration dependence of expansion speeds in glucose only (open green squares), glycerol only (open black circles), glycerol or glucose with aspartate (red circles, purple squares); data for latter same as shown in Fig. 4d and Extended Data Fig. 7e, f. Expansion speed in glucose ($\sim 1\text{--}2\text{ mm h}^{-1}$) is faster than in glycerol (not an attractant) but well below the cases for which (low) amounts of attractants are supplemented. Shown data points represent means of biological replicates ($n = 2$ or larger), with error bars (s.d.) shown for $n \geq 3$; see Supplementary Tables 9, 10 for data and sample sizes. **d**, **e**, To understand the expansion behaviour, we used confocal scans to obtain the density profiles. The ring observed in the photograph is seen as a subtle density bulge at the front bounding a flat-density interior. Note the lack of an exponential trailing region, as observed when an attractant supplement is present (Fig. 2b, Extended Data Fig. 4i, photographs in Fig. 1a). The observed density profiles are comparable with those previously found with galactose as the attractant and the major nutrient source¹⁸. Experiments here were done with wild-type cells (HE206) (**a–c**) and fluorescence cells (HE274) (**d**, **e**). The confocal experiments were conducted once (expansion speeds are highly comparable to those measured manually). **f**, To capture the observed

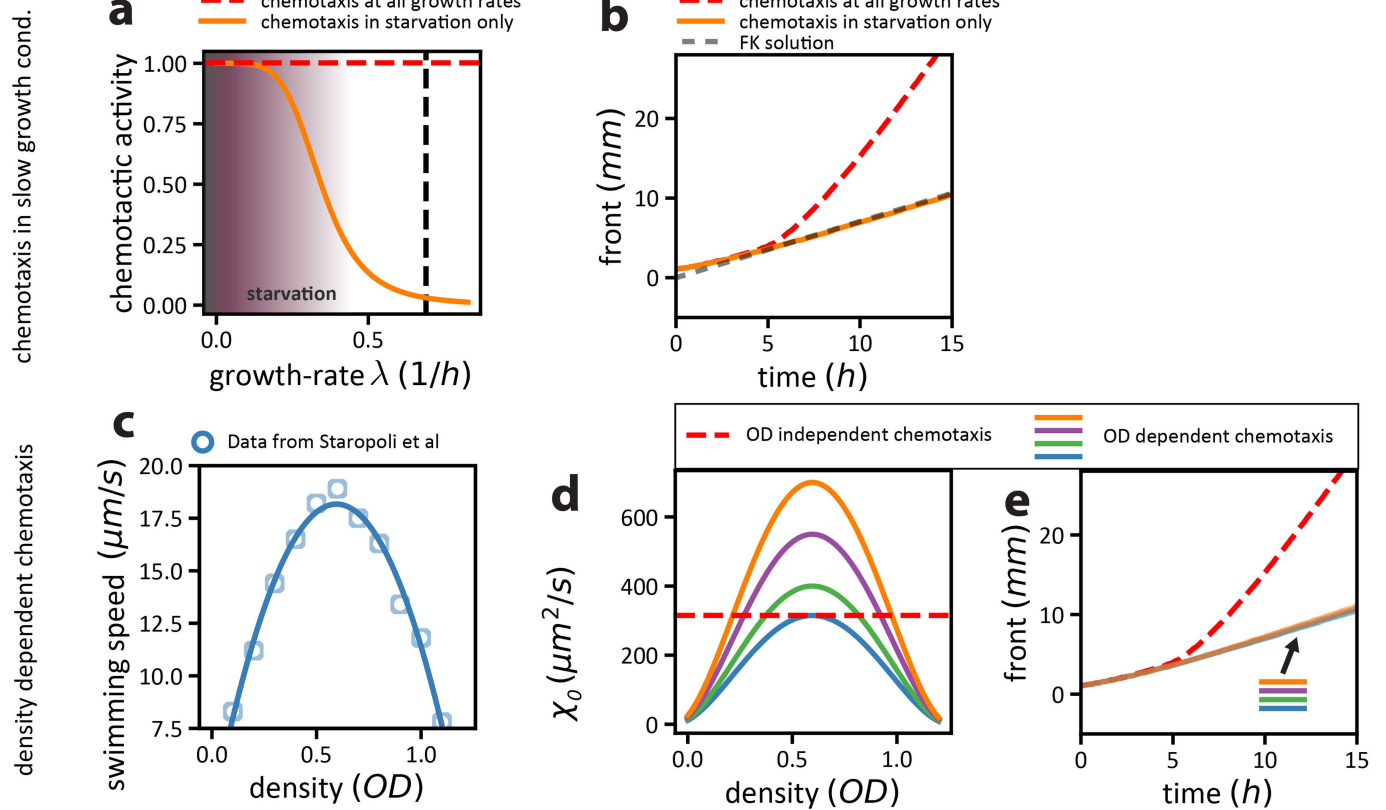
behaviours, we modified the GE model (Fig. 3a) using only one variable a to describe the attractant/nutrient. Consumption of the growth-enabling attractant is directly coupled to the increase in density via the yield Y . **g**, **h**, Fixing model parameters using available data for growth and chemotaxis on glucose (see Supplementary Text 2.4, 2.3 with parameters used listed in Supplementary Table 4), the model generated expansion speeds (green line in **b**) and density profiles that capture the experimental observations well; for comparison, a coarse-grained spatial resolution similar to the experiments was used to display the profiles obtained by the simulations. **i**, The model output can further be understood by a scaling analysis (Supplementary Text 2.6), resulting in the simple relation $u^2 \propto \chi_0 \lambda$ (equation (E8.e)). This relation is of the same form as the result of the Fisher–Kolmogorov dynamics, $u_{\text{FK}} = 2\sqrt{D\lambda}$ (see Extended Data Fig. 5), but with the chemotactic coefficient χ_0 replacing the diffusion coefficient D . **j**, **k**, The predicted dependence of u on λ and χ_0 (black lines) is validated by numerical simulations of the model (blue circles). The square-root dependence of the expansion speed on the chemotactic coefficient χ_0 stands in contrast to the linear dependence on χ_0 when an attractant supplement is provided (Extended Data Fig. 9d) and shows that the expansion dynamics with or without the attractant supplement are two distinct classes of mathematical problem. Note that the quantitative gain in expansion speed for the case with a supplemented attractant comes not only from the change in dependence on the chemotactic coefficient from $\sqrt{\chi_0}$ to χ_0 , but also from the freedom to use attractants that have large χ_0 but small λ , which can be compensated by nutrients that give larger λ . Both aspartate and serine are strong attractants but poor nutrients, and are thus most potent when used in combination with a good nutrient source. Thus, separating the role of substances as nutrients and as cues not only relaxes the underlying mathematical constraint but also relaxes the biological constraint so that good attractants need not be good nutrients. These results provide an important support for the central thesis of this work, that chemotactic cells gain fitness by expanding in nutrient-replete conditions as a ‘foresighted’ navigation strategy (see main text).



Extended Data Fig. 9 | Scaling analysis of expansion dynamics and

illustration of the stochastic migration process. **a**, The exponential trailing region of the density profile is fixed by the cell growth rate λ and expansion speed u of the front. Because cells in the trailing region do not experience drift (Extended Data Fig. 3g), the apparent ‘movement’ of the trailing region at the same speed as the front bulge is possible only if it has an exponential profile, $\rho(r, t) \sim e^{k(r - ut)}$, with $k = \lambda/u$. **b**, Scaling of the expansion speed with model parameters. According to the GM model (Fig. 3a), the density peak at the propagating front is determined by a balance between cell growth and back diffusion (Fig. 4b). Using a crude scaling analysis to capture this balance, we can obtain (approximately) the quantitative determinants of the propagating speed. Consider a sharply peaked density bulge at the front, with peak density ρ_{peak} and width w . The number of cells contained in the peak region, N_{peak} , is given by the relation in equation (E9.a). Cell birth rate, λN_{peak} , is balanced by the back-diffusion flux, which is approximated as $D \rho_{peak} / w$, leading to the relation in equation (E9.b). To relate to the migration speed u , we note that around the density peak the drift speed v is nearly maximal (Fig. 4a), and equation (4) becomes $v_{max} \approx \chi_0 \frac{d}{dx} \ln(a)$. In the scaling approach, we take $u \sim v_{max}$ and the approximation $\frac{d}{dx} \ln(a) \sim 1/w$ leading to the relation in equation (E9.c).

Combining equations (E9.a) and (E9.c), we obtain equation (E9.d) with the expansion speed increasing with the square-root of the growth rate λ . Note the χ_0/D factor appearing as a prefactor in the expression for u , which is responsible for the increase in the expansion speed in the presence of chemotaxis with respect to the Fisher–Kolmogorov dynamics (Extended Data Fig. 5) and for the dynamics with the attractant being the sole nutrient (Extended Data Fig. 8). **c, d**, Scaling results are confirmed by simulations of the GE model when varying growth rate λ (**c**) and chemotactic coefficient χ_0 (**d**). **e**, To further illustrate the intricate dynamics at the front of the expanding population, we performed stochastic agent-based simulations looking at the trajectories of single cells. Shown here are cell trajectories for a few selected cells located within the population front (pioneers) at time $t = 6.5$ h. Bottom, 38 trajectories with colour indicating the time the trajectory escaped from the front and cells switched from being pioneers to being settlers, which grow and colonize localities behind the front. Red circles indicate cell division events. Highlighted area (cyan) denotes front region with aspartate concentration in the range $a_- < a < a_+$. Top, position distribution of all simulated trajectories (1,000) at time $t = 8$ h. See Supplementary Text 3 for details.



Extended Data Fig. 10 | Modelled scenario of chemotaxis as a strict starvation response. To examine the expansion characteristics of the population under the hypothetical scenario in which chemotaxis is a strict starvation response, we modified the GE model (Fig. 3a) to investigate the cases when chemotaxis is active either only in slow growth conditions (**a, b**) or within intermediate density ranges (**c–e**). **a**, To model chemotaxis being activated at slow growth, we introduced a strong dependence of chemotaxis on local growth rate (orange line). In contrast to the original GE model (dashed red line), we used a chemotactic coefficient that depends on growth conditions, $\chi_0 = \chi_0(\lambda(n))$ (orange line). Black dashed line shows growth rate in the presence of saturating glycerol. **b**, This dependence of chemotaxis on growth conditions leads to a marked decrease in the speed of expansion (compare orange and red dashed lines). The expansion dynamics of this model resembles the Fisher–Kolmogorov dynamics (grey dashed line), suggesting that chemotaxis does not boost population-level expansion when it is activated only under slow growth conditions. **c–e**, We further studied the case of swimming being a density-dependent response, active only at intermediate

bacterial densities, as has been observed in batch culture measurements¹⁵ (**c**). Taking such a dependence of the swimming speed (v) on the local cell density (ρ) and assuming $\chi_0 \sim v^2(\rho)$, we looked at the expansion dynamics for several maximum values of the chemotactic coefficient (**d, e**). For all of the forms of $\chi_0(\rho)$ shown in **d**, population expansion was slowed down substantially as compared to the reference case in which chemotaxis is also active at low densities (red dashed lines). The slow expansion dynamics is again similar to the Fisher–Kolmogorov dynamics, illustrating that the boost of expansion speed and population size by chemotaxis relies on chemotaxis being active at low densities. Note that in both cases analysed, we have not included the dependence of the diffusion constant on growth rate or local densities but assumed a constant value as in the original GE model. Introducing such dependences would further reduce the speed of expansion, below even that of Fisher–Kolmogorov dynamics. The origin of the slow expansion dynamics in these models is simple: a population cannot expand faster than its front, and the front is at low density and experiences the fastest growth rate.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Microscopic imaging data were collected and analyzed using the software LAS AF SP8 (Leica Microsystems) and a custom-made Python script (Python 2.7) available via GitHub (see code availability statement).
Data analysis	Custom made scripts were used to analyze experimental data and to perform simulations (full description in supplementary text). Scripts are available from the authors upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Major experimental data supporting this study are provided in this manuscript or available via figshare repositories: doi.org/10.6084/m9.figshare.9639209 (confocal expansion data) and doi.org/10.6084/m9.figshare.9643001 (data swimming observation). Simulation data can be generated with the provided simulation code and parameter sets.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Major quantities (growth-rate, expansion speed, swimming speed, and uptake-rates) in reference conditions confirmed small run to run variations (<10%, based on 3-9 biological replicates). Based on this finding we repeated most of the measurements shown in this study at least once (n=2 biological replicates). All repeats confirmed similar results. Confocal microscopy observations were conducted twice for the reference condition (glycerol+aspartate) showing again similar results. Microscopy observation were conducted once for other conditions.
Data exclusions	No data were excluded from the analysis.
Replication	All biological replicates showed comparable results.
Randomization	Not applicable.
Blinding	Data collection followed the same predetermined protocols throughout the whole study; the analysis of swimming behavior and imaging data is based on custom made codes without adjustable parameters. Blinding was therefore not used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

An evolutionarily stable strategy to colonize spatially extended habitats

<https://doi.org/10.1038/s41586-019-1734-x>

Weirong Liu^{1,2,5}, Jonas Cremer^{3,4,5}, Dengjin Li^{1,2}, Terence Hwa^{3*} & Chenli Liu^{1,2*}

Received: 12 September 2018

Accepted: 3 October 2019

Published online: 6 November 2019

The ability of a species to colonize newly available habitats is crucial to its overall fitness^{1–3}. In general, motility and fast expansion are expected to be beneficial for colonization and hence for the fitness of an organism^{4–7}. Here we apply an evolution protocol to investigate phenotypical requirements for colonizing habitats of different sizes during range expansion by chemotaxis bacteria⁸. Contrary to the intuitive expectation that faster is better, we show that there is an optimal expansion speed for a given habitat size. Our analysis showed that this effect arises from interactions among pioneering cells at the front of the expanding population, and revealed a simple, evolutionarily stable strategy for colonizing a habitat of a specific size: to expand at a speed given by the product of the growth rate and the habitat size. These results illustrate stability-to-invasion as a powerful principle for the selection of phenotypes in complex ecological processes.

When an organism encounters an unoccupied habitat, it colonizes the habitat through growth and expansion^{2,3,8–13}. A recent quantitative study of chemotaxis-mediated bacterial range expansion¹⁴ showed that the characteristics of the expanding population are dominated by a group of pioneering cells at the population front; these pioneers move outwards, replicate, and leave behind offspring (settlers) to grow and occupy the territories traversed (Extended Data Fig. 1a, b, Supplementary Video 1). To understand the determinants of colonization behind the front, we modified common experimental evolution protocols^{4–7,15,16} to select for cells at various distances from the point of the initial invasion, well after the passage of the front. Motile *Escherichia coli* cells inoculated at the centre of a semi-solid tryptone broth (TB) agar plate were given time to expand outwards and colonize the entire plate^{8,14,17}. Twenty-four hours after inoculation, after the entire plate had been filled with bacteria, a small volume of agar (containing about 7.4×10^6 cells) was taken at a specific radius and transferred directly to the centre of a fresh plate (Fig. 1a). The process was repeated for 50 cycles (about 600 generations), with the transferred samples always taken at the same distance from the centre. Five such series were generated at five different distances from the origin (positions A–E in Extended Data Fig. 1c).

We first evaluated the growth and expansion characteristics of samples collected at various cycles and positions. When inoculated on fresh TB plates, the evolved populations still expanded steadily outwards (Extended Data Fig. 1d), but were characterized by expansion speeds with distinct dependencies on the selection distance (Fig. 1b). Populations collected from outer radii exhibited a steady increase in expansion speeds, whereas those collected from inner radii exhibited a steady decrease, leaving an intermediate selection distance (position C, with distance $X_C = 15$ mm) at which the expansion speed of the evolved populations fluctuated around that of the ancestor (about 6 mm h^{-1}) throughout the evolution process. These results were highly

reproducible across replicates and in different growth media containing generic amino acid supplements (Extended Data Fig. 2a–c). The divergent pattern of the evolved expansion speeds obtained (Fig. 1b) is not the result of a simple trade-off with the rate of cell growth^{4,5,18}, as the batch culture growth rates changed very little for strains evolved in TB (Extended Data Fig. 1e). Strains evolved in casamino acids (CAA) generally showed growth rates higher than the ancestor (Extended Data Fig. 1f, g); however, changes in their expansion speeds still followed the divergent pattern according to their selection positions (Extended Data Fig. 2c). By contrast, expansion speeds increased regardless of selection position for strains evolved in medium with glycerol and no chemoattractants (Extended Data Fig. 2d), suggesting that chemotaxis is important for the divergent evolution phenomenon shown in Fig. 1.

An examination of 300 strains sampled from the 50th cycle of the five evolved series across three replicates showed that the distributions of expansion speeds of individual strains were well reflected by the previous measurements of samples containing mixed populations (Fig. 1c). Furthermore, changes in expansion speeds were consistent with changes in the motility characteristics of the evolved cells obtained from single-cell analysis (Extended Data Fig. 1h, i), and with mutations identified from genomic sequence analysis. Sequencing of population samples at the 50th cycle of each evolved series yielded a multitude of mutations (Supplementary Table 1). Several dominant mutations were introduced individually into the ancestral strain; these were found to change the expansion speeds of the ancestral strain towards those of the evolved strains from which the mutations were derived (Extended Data Fig. 1j).

Two-strain competition in space

To understand the underlying evolutionary process, we first compared the expansion dynamics of clonal populations that were grown

¹CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China. ²University of Chinese Academy of Sciences, Beijing, People's Republic of China. ³Department of Physics, University of California San Diego, La Jolla, CA, USA.

⁴Present address: Department of Molecular Immunology and Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands. ⁵These authors contributed equally: Weirong Liu, Jonas Cremer. *e-mail: hwa@ucsd.edu; clliu@siat.ac.cn

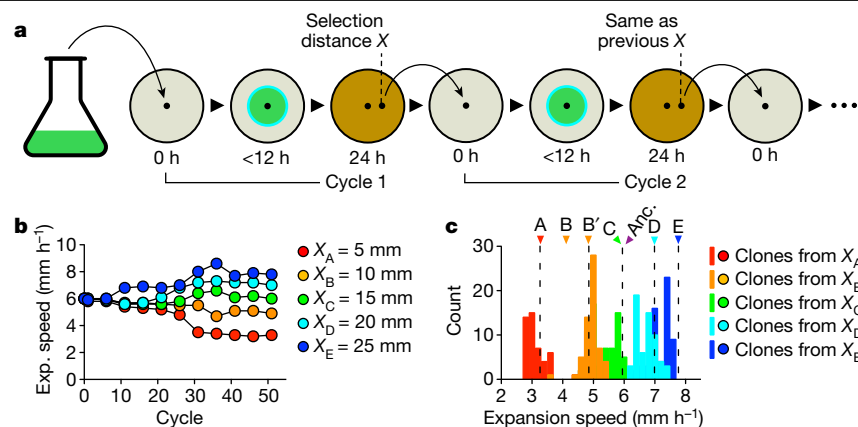


Fig. 1 | Experimental evolution with position-dependent selection.

a, Schematic illustration of the evolution experiment. Exponentially growing *E. coli* CLM cells²⁸ (ancestor) were inoculated at the centre of a 0.25% TB agar plate; after 24 h incubation at 37 °C, when the bacteria have colonized the entire plate, a 2- μ l cell–agar mixture was taken at a distance X away from the plate centre and inoculated at the centre of a fresh plate (see Methods). The cycle was repeated with the samples always taken after 24 h at the same distance X as in previous cycles. **b**, Expansion speeds of population samples from each of the five selection series are shown from lineage 1 at various evolution cycles.

Experiments were repeated three times with similar results (see text and Extended Data Fig. 2). **c**, Expansion speeds of 300 individual clones (60 for each selection series) isolated from the population samples at the 50th cycle. A histogram for each series is shown with the corresponding colours. The average expansion speeds of the mixed population of each selection series at the 50th cycle (Extended Data Fig. 1d) are shown as dashed lines for comparison. The expansion speeds of the ancestor and six mutant strains isolated at the 50th cycle (strains A, B, B', C, D, E) are indicated at the top; these resemble the average expansion speed of the corresponding mixed population of lineage 1.

individually. We chose several mutant strains isolated at the 50th cycle that exhibited a range of expansion speeds but had similar growth rates (A, B, B', C, D, and E in Fig. 1c; see Supplementary Table 2). We transformed each strain with GFP and calibrated their fluorescence intensities by direct cell counting (Extended Data Fig. 3a–c). This allowed direct observations of the spatiotemporal dynamics of density profiles of each strain (Extended Data Fig. 4a–c for ancestor, mutant B, and mutant D). Clearly, faster strains showed higher abundances at all positions and all times.

Next, we competed each mutant strain against the ancestor strain. We transformed these strains with a non-fluorescent variant of GFP and ensured that each had a similar growth rate and expansion speed to the same strains with fluorescent GFP (Extended Data Fig. 3a–f). Equal mixtures of various strain pairs were inoculated at the centres of agar plates and their spatiotemporal abundance patterns were characterized (see Methods). Figure 2a shows the outcome of competition between mutant D and the ancestor 12 h after inoculation. Notably, the two strains dominated different spatial regions: the ancestor (pseudocoloured purple) dominated the interior whereas mutant D (pseudocoloured cyan) dominated the exterior. Repeating the competition process between mutant B and the ancestor (Fig. 2b), we found the opposite, with strain B dominating the interior and the ancestor dominating the exterior. The ratio of the calibrated fluorescence intensities, shown as the coloured solid lines in Fig. 2c, d, agree well with the ratio W of the densities of evolved cells over the ancestor (circles in Fig. 2c, d) as obtained from cell counting (Extended Data Fig. 4f), a direct measure of the relative fitness of the mutant¹⁹ at each location. This relative fitness profile was stable through much of the 24-h course of competition (Extended Data Fig. 4g–j).

As strain B expanded slower than the ancestor whereas strain D expanded faster (Fig. 1c), the competition results suggest a trend in which the slower strain dominates the interior and the faster strain dominates the exterior. This is in stark contrast to the ratio of cell densities from single-strain expansion dynamics (coloured dashed lines in Fig. 2c, d, derived from Extended Data Fig. 4b), which shows an advantage for the faster strain everywhere. Thus, the faster strain became disadvantaged in the interior only when grown in the presence of the slower strain, manifesting the ‘game-like’ nature of the underlying evolutionary process^{20,21}; that is, the fitness of a strain at a location depends on the presence and motility of competing strains.

Repeating the competition assay between the ancestor and each of the six mutants, we found that the slower strain dominates inside and the faster strain outside in each case (Extended Data Fig. 4k–m). The competition results can be concisely summarized by defining a crossover distance d_x at which the ancestor and the mutant have the same fitness: $W(d_x) = 0$. This is indicated as the vertical dashed line in Fig. 2c, d and Extended Data Fig. 4m. This crossover distance is plotted against the expansion speed of the corresponding mutant in Fig. 2e, with various regions shaded according to strain dominance: the faster strain for $X > d_x$ and the slower strain for $X < d_x$. To see whether the competition results represented by this ‘phase diagram’ are specific to the evolved strains, we repeated these studies using two synthetic strains, WL1 and WL2 (Supplementary Table 2), which allowed us to titrate swimming speed and expansion speed by using specific inducers without affecting cell growth (Extended Data Fig. 3g–l). When we competed the fluorescent versions of these strains with each other, we obtained a phase diagram (Extended Data Fig. 3m) that was very similar to that between the ancestor and evolved strains (Fig. 2e). This indicates that the latter represents a generic outcome of competition between strains with different motility characteristics, regardless of how these characteristics are changed.

Modelling competitive expansion dynamics

To gain more insight into the competition dynamics, we turned to a mathematical model of bacterial population expansion¹⁴ that includes the effect of cell growth along with the random and directed components of cell motion, based on well-characterized molecular interactions^{14,22–27} (Extended Data Fig. 5a). This model provides a quantitatively accurate description of the expansion dynamics for a single bacterial strain in soft agar¹⁴ (Extended Data Fig. 5b, c). We extended this model to describe competition between two strains that are assumed to respond to the same chemoattractant and grow at the same rate (Extended Data Fig. 6a), with different expansion speeds modelled by different parameters characterizing chemotaxis (see Supplementary Model for details and Supplementary Video 2 for an example of the dynamics). This model captured the spatial dominance pattern of the slow and fast strains observed after a long time (Extended Data Fig. 6b), as well as the time dependence of the crossover distance d_x (Extended Data Fig. 6c). Factors that favour the dominance of slower strains at smaller

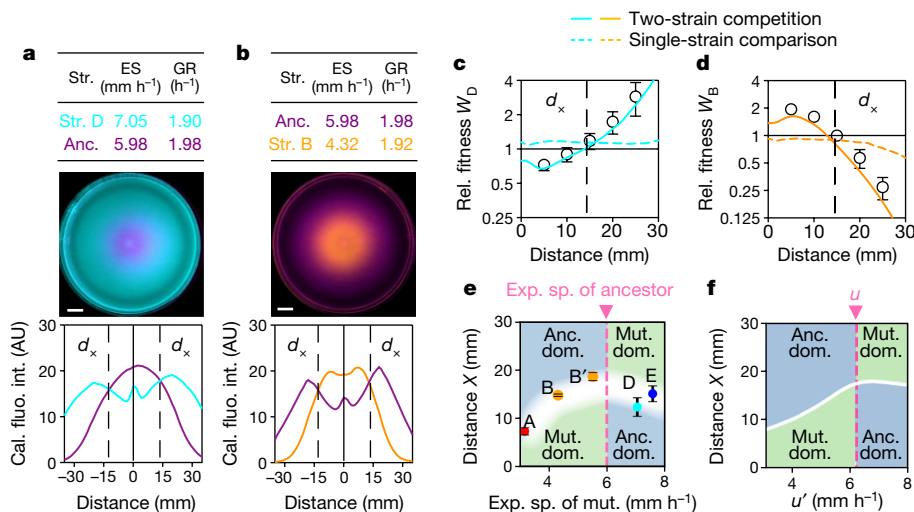


Fig. 2 | Competitive expansion in space. **a, b**, Expansion speeds and growth rates (top), merged pseudocolour images (middle) and cell density profiles (bottom) of representative two-strain competitions between the fluorescent derivatives (Extended Data Fig. 3) of the ancestor (anc.; purple) and mutant strains (str.) D (**a**, cyan) and B (**b**, orange). Cal. fluo. int., calibrated fluorescence intensity. See Extended Data Fig. 4d, e for raw images and fluorescence intensity profiles before merging, and Methods for details of the competition assay. Strain details are given in Supplementary Table 2. Scale bars, 10 mm. The data were taken 12 h after co-inoculation of equal initial mixtures of the two competing strains at the centre of 0.25% TB agar plates. Data show means of three biological replicates. **c, d**, The relative fitness W_i of strains $i = \{D, B\}$ (relative to the ancestor), defined here as the ratio of cell densities at different positions. Coloured solid lines were obtained as the ratio of the fluorescence density profiles shown in Extended Data Fig. 4d, e (bottom). Open circles indicate ratios of direct cell counts (Extended Data Fig. 4f). Coloured dashed lines indicate the ratio of fluorescence between the mutant and ancestor, when each was grown individually in the same agar plate for 12 h after inoculation at the centre; results derived from data shown in

Extended Data Fig. 4b. **e**, Circles indicate crossover distances for the competition of the ancestor with five evolved strains (Extended Data Fig. 4k–m), plotted against their respective expansion speeds (exp. sp.) (Extended Data Fig. 3b). Dashed vertical line indicates the expansion speed of the ancestor. For the competition of strain C with the ancestor, their expansion speeds were very close and the relative fitness was difficult to resolve (Extended Data Figs. 3b, 4k–m). The background colour for regions above and below the crossover distances indicates distances at which the ancestor (blue) or mutant (mut.; green) dominates. **f**, The white line shows the crossover distance according to the competitive expansion model (Supplementary Model) for two strains with expansion speeds u and u' , with u ('ancestor' speed) fixed (dashed vertical line) and u' (competing 'mutant' speed) varied. The background colours again indicate the regions of dominance by either strain. The regions of strain dominance are assigned according to the simple rule that the faster strain dominates for $X > d_x$ and the slower strain dominates for $X < d_x$. Experiments in **a, b** were repeated independently three times with similar results. For **c–e**, data are mean \pm s.d. for $n = 3$ biological replicates.

distances are attributed to shifts in balance among cell growth, forward movement, and back-propagation of pioneering cells¹⁴ from the population front (Extended Data Fig. 7, Supplementary Analysis 1). Using this competitive expansion model, we systematically computed the outcome of competition between an equal initial mixture of two strains, varying the expansion speed of one strain (the mutant, with speed u') while holding that of the other (ancestor strain, with speed u) at a fixed value. Figure 2f shows the crossover distances $d_x(u, u')$ and the resulting phase diagram obtained for these competitions, which are similar to those observed experimentally (Fig. 2e, Extended Data Fig. 3m).

We also used the competitive expansion model to probe the dependence of spatial dominance for three strains. Consider three strains (**a, b, c**) with single-strain expansion speeds u_a, u_b, u_c such that $u_a < u_b < u_c$. Let us find the region of dominance by strain **b** during three-strain competition. From the two-strain crossover distances $d_a = d_x(u_b, u_a)$ and $d_c = d_x(u_b, u_c)$ between strains **a–b** and **b–c**, respectively, the illustration in Fig. 3a clearly suggests that strain **b** will dominate in the region $d_a < d < d_c$. This is verified experimentally by directly competing three strains with different expansion speeds (that is, strains **A, C, E** in Fig. 3b). Thus the outcome of three-strain competition can be correctly predicted from the results of two pairwise two-strain competitions (in particular, the form of the crossover distance). We next show that this simple result can be generalized to predict the outcomes of the evolution experiments shown in Fig. 1.

An evolutionarily stable strategy

To connect to these evolution experiments, which involve potentially many strains with a continuum of expansion speeds, let us first consider

the theoretical limit $u_a \rightarrow u_b$ and $u_c \rightarrow u_b$ (black arrows in Fig. 3a). In this case, the region where strain **b** dominates will be pinched and distributed narrowly around a special distance, $d_b = \lim_{u' \rightarrow u_b} d_x(u_b, u')$, the distance at which the strain with speed u_b is dominant over other strains even if their speeds are infinitesimally different. As there is nothing special about strain **b** and its speed, this consideration suggests a much more general result: that for a strain with a single-strain expansion speed u , there is a special distance

$$d^*(u) \equiv \lim_{u' \rightarrow u} d_x(u, u') \quad (1)$$

at which no other strain with a different speed can dominate.

Given the form of the crossover distance shown in Fig. 4a, the diagonal $d^*(u)$ (pink dashed line) as defined by equation (1) turns out to depend linearly on u as shown in Fig. 4b. This simple result is reinforced by a more detailed mathematical analysis (Supplementary Analysis 1, 2), which further predicts that the slope of d^* versus u will be inversely proportional to the growth rate λ :

$$d^*(u) \propto u/\lambda \quad (2)$$

This form is confirmed by numerical simulation of the competitive expansion model performed at different growth rates (Extended Data Fig. 6d, e).

So far, equations (1) and (2) refer to the dominance of a strain in a 50:50 initial mixture of it with a competing strain. To apply these results on multi-strain competition to evolutionary dynamics, in which mutants may be generated at very low frequencies, it is necessary to recalculate the crossover distance $d_x(u, u')$ for a low frequency of the competing

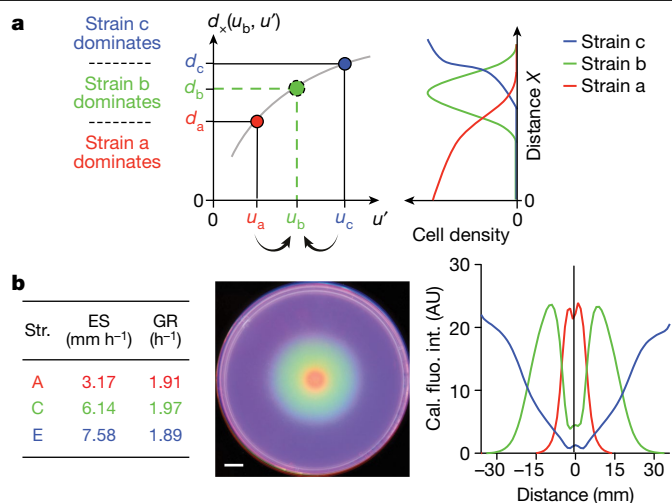


Fig. 3 | Three-strain competition and stability distance. **a**, The crossover distance $d_x(u, u')$ for two competing strains with speeds u and u' encode important information regarding the competitive expansion dynamics. We illustrate this by considering three strains (a, b, c) with single-strain expansion speeds u_a, u_b, u_c . To see the region of dominance by strain b, we sketch the crossover distance $d_x(u, u')$ versus u' for $u = u_b$. The red point indicates the crossover distance d_a between strains a and b, with $d_a = d_x(u_b, u' = u_a)$. Similarly, the blue point indicates the crossover distance $d_c = d_x(u_b, u' = u_c)$ between strains b and c. By the definition of the crossover distance, strain a (which is slower than b) would dominate at distances $d < d_a$ and strain c (which is faster than b) would dominate at distances $d > d_c$. Hence there is a region $d_a < d < d_c$ where strain b dominates. In the limit $u_a \rightarrow u_b$ and $u_c \rightarrow u_b$, the regime of dominance by strain b becomes narrowly distributed around $d_b = \lim_{u' \rightarrow u_b} d_x(u_b, u')$: the green point. The analysis described here can be performed more rigorously to formulate an evolutionary stability criterion (Extended Data Fig. 8). **b**, Expansion speeds (ES) and growth rates (GR) (left), merged pseudocolour images (middle) and cell density profiles (right) of representative three-strain competitions between the fluorescent derivatives of strains A (red), C (green), and E (blue). Strain details are given in Supplementary Table 2. Scale bar, 10 mm. Data were taken 12 h after co-inoculation of equal initial mixtures of the three competing strains at the centres of 0.25% TB agar plates; repeated independently three times with similar results.

strain. However, as we show in Extended Data Fig. 8, for two strains with comparable expansion speeds, their crossover distance is independent of the frequency of the competitor. Thus, equation (2) can be applied to competitions involving strains of arbitrarily small frequencies, including spontaneously generated mutants. Therefore, equations (1) and (2) describe an evolutionary stability criterion, that a strain with expansion speed u is stable against invasion by mutants with different expansion speeds at position $d^*(u)$ as given by equation (2). We therefore refer to d^* as the stability distance and $d^*(u)$ as the stability line.

The actual selection experiments performed (Fig. 1) pose a slightly different question from the evolutionary stability criterion just described: at a given selection distance X , what speed $u^*(X)$ is most fit? The answer is just the mathematical inversion of equation (2):

$$u^*(X) \propto X\lambda \quad (3)$$

This result can be appreciated by examining a slice of the crossover landscape of Fig. 4a, for $d_x(u, u') = X$ as shown in Fig. 4c. The stable speed selected is at the intersection of $d_x(u, u') = X$ (cyan line) and the diagonal ($u = u'$, pink line), indicated by the circle, as a strain with speed that deviates from the intersection (black arrows) is selected against; see legend for details. In the plot of the stability line (Fig. 4b), we added the teal-coloured secondary axes: at a given distance X , the selected speed $u^*(X)$ is obtained by following the teal arrows, whereas for a strain with a given speed u , its stability distance is obtained by following the black arrows.

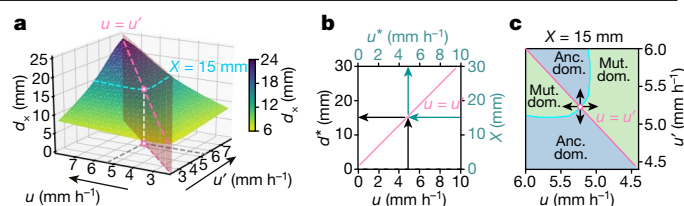


Fig. 4 | Stability line of the competition dynamics. Competition results for pairwise combinations of expansion speeds (u, u') following the growth-expansion model¹⁴ (Extended Data Figs. 5, 6). **a**, The crossover distances $d_x(u, u')$ at which both strains are equally abundant (green surface). The considerations in Fig. 3a indicate that the diagonal (pink dashed line where $u = u'$) gives the stability distance d^* of a strain with ancestral speed u . **b**, The relation $d^*(u)$ is linear and is called the stability line. This line has an orthogonal interpretation: following the teal arrows, it gives the expansion speed u^* that would be selected at position X . To verify this orthogonal view, we note that at a given distance, for example, $X = 15$ mm, there is a set of expansion speed combinations for which the crossover distances correspond to this distance ($d_x = X$, cyan line in **a**). A distinct speed u^* among this set is indicated in white, corresponding to the limiting value of d_x when the expansion speeds of the two competing strains approach each other, defined mathematically from $X = \lim_{u, u' \rightarrow u^*} d_x(u, u')$. **c**, A strain with this special expansion speed $u^*(X)$ is stable against mutants with different speeds at distance X , according to the strain dominance pattern shown: different regions in this panel are assigned the same way as in Fig. 2f. For $u > u^*$, the 'ancestor strain' dominates where the green surface in **a** is below X , and the 'mutant' dominates where the green surface is above X ; vice versa for $u < u^*$. The expansion speed $u^*(X)$ is located where the phase boundary (cyan line) intersects the diagonal. Here, if the speed of one strain is increased or decreased (arrows), then it is selected against as the other strain dominates.

Validation of the stability criterion

To test the predicted stability line (equations (2), (3)), we designed two additional sets of evolution experiments for growth conditions that provided altered ancestral expansion speed and growth rate. First, we repeated the evolution protocol shown in Fig. 1a using the same growth medium (TB) but different agar densities. This changed the effective cell diffusion constant²⁸, resulting in a range of expansion speeds (purple squares, Fig. 5a) without affecting the growth rate¹⁴. The evolution results obtained for five agar densities, each for five selection distances and different replicates, were highly reproducible (Extended Data Fig. 9a). From the evolved expansion speeds, we determined the stable selection distance for each agar density (Extended Data Fig. 9b, c). The stability distances obtained exhibited a linear relation with the ancestral expansion speeds, as predicted (Fig. 5b).

Next we plotted the expansion speeds obtained from different cycles of evolution (the data in Extended Data Fig. 9a) with the stability line in Fig. 5c–e for data from three different agar concentrations. Interpreting the stability line as the stable expansion speed $u^*(X)$ at the corresponding selection distance X (Fig. 4b), the data from each evolution series (symbols of the same colour) are seen to approach the predicted final stable values.

Then we repeated all of the evolution experiments yet again, at various selection distances and agar densities, in a medium that supports approximately 50% slower growth (CAA; brown squares in Extended Data Fig. 9d). The expansion speeds of the evolved strains (Extended Data Fig. 9e) exhibited a similar pattern of changes to those obtained in TB (Extended Data Fig. 9a). The stability distances obtained for different ancestral expansion speeds (Fig. 5f) again followed a linear relation as predicted, with an approximately 70% increase in the slope, consistent with the dependence on cell growth rate given in equation (2).

To probe the generality of a stable expansion speed and its linear size-dependence, we investigated another mode of selection *in silico* using the multi-strain generalization of the competitive expansion model (Supplementary Analysis 5). In this mode of selection, a fraction

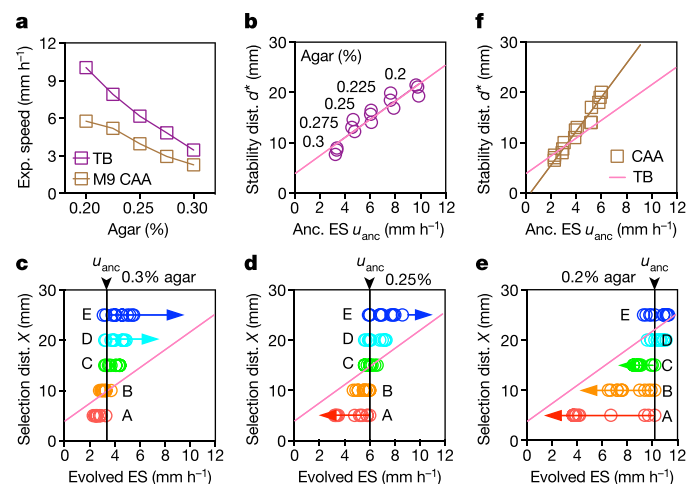


Fig. 5 | Validation of the evolutionary stability line. **a**, Expansion speed of the ancestral strain can be modified by changing the agar density in the range 0.2–0.3%. Results are shown for both TB and CAA as the nutrient source. **b**, Estimated stability distance d^* obtained as in Extended Data Fig. 9b, c at different agar densities; results for each of the three replicates are shown. Pink line, linear fit. **c–e**, Expansion speeds from every five cycles of evolution (Fig. 1b) are plotted at each selection distance (X_A, \dots, X_E) for agar density of 0.3% (**c**), 0.25% (**d**) and 0.2% (**e**). The pink line is the expected stable attractor of evolution dynamics obtained in **b**. **f**, Stability distances estimated for each agar density in CAA, for each of the three replicates. Brown line, linear fit. The slope is steeper than the pink line (from **b** for TB), as predicted for the slower growth rate in CAA. In **a**, means for $n = 3$ biologically independent repeats are shown (s.d. error bars are smaller than the symbols). Experiments in **c–e** were repeated independently three times with similar results (Extended Data Fig. 9).

of all cells contained within a habitat of a certain size were repeatedly propagated to the next cycle; see Extended Data Fig. 10a. After just a few cycles of simulations, the average expansion speeds of the populations diverged according to habitat sizes, with smaller or larger habitats dominated by species with slower or faster expansion speeds, respectively (Extended Data Fig. 10b). This follows the trend seen in our evolution experiment with selection at a fixed distance (Fig. 1), despite two-dimensional weighting and edge effects, which made the slower species take more cycles to dominate in the smaller habitat (Extended Data Fig. 10c, d). Thus, the results of the spatial competition and evolution scenario studied here are robust to specific implementations of the selection process, and may be applicable to populations confined to environments with finite resource patches (for example, nutrient hotspots or restricted physical terrains).

Fitness effects that depend on the composition of the population are prevalent in natural evolution^{29–31}. The complex dynamics that result from these effects were investigated theoretically early on in the form of simple evolutionary games^{32,33}; recent experimental studies have also considered the consequences of composition-dependent reproduction in synthetic microbial populations^{12,34}. Here we encounter a natural example in which the dominance of one strain at a location depends not only on its own expansion speed but also on the expansion speed of another strain (Figs. 2c, d, 3b, Extended Data Fig. 3m) as well as on the initial abundance of the other strain (Extended Data Fig. 8). Dominance patterns such as that seen in Fig. 4c are analogous to the ‘payoff matrix’ of a game, with the expansion speed being a strategy taken by a player^{12,20,21,33,34}; the cyan stability line can be viewed as the set of coexistence points of this game. Notably, such a highly complex game involving spatial selection and the composition dependence of fitness can be elucidated in quantitative details, with quantitative predictions of the selected phenotypes in different environments, by simply identifying the stable equilibria of the complex dynamics. As there is no fixed fitness landscape to ‘climb’, the winners of this

evolutionary process are not the fittest in an absolute sense, but simply those that are stable against invasion by mutants. This provides a fresh perspective for approaching other complex evolutionary problems in nature, where composition-dependent effects are ubiquitous.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1734-x>.

- Hanski, I. Metapopulation dynamics. *Nature* **396**, 41–49 (1998).
- Skellam, J. G. Random dispersal in theoretical populations. *Biometrika* **38**, 196–218 (1951).
- Andow, D. A., Kareiva, P. M., Levin, S. A. & Okubo, A. Spread of invading organisms. *Landsc. Ecol.* **4**, 177–188 (1990).
- Yi, X. & Dean, A. M. Phenotypic plasticity as an adaptation to a functional trade-off. *eLife* **5**, e19307 (2016).
- Fraebel, D. T. et al. Environment determines evolutionary trajectory in a constrained phenotypic space. *eLife* **6**, e24669 (2017).
- Ni, B. et al. Evolutionary remodeling of bacterial motility checkpoint control. *Cell Rep.* **18**, 866–877 (2017).
- Shih, H.-Y., Mickalide, H., Fraebel, D. T., Goldenfeld, N. & Kuehn, S. Biophysical constraints determine the selection of phenotypic fluctuations during directed evolution. *Phys. Biol.* **15**, 065003 (2018).
- Adler, J. Chemotaxis in bacteria. *Science* **153**, 708–716 (1966).
- Levin, S. A. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* **73**, 1943–1967 (1992).
- Hastings, A. et al. The spatial spread of invasions: new developments in theory and evidence. *Ecol. Lett.* **8**, 91–101 (2005).
- Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl Acad. Sci. USA* **104**, 19926–19930 (2007).
- Müller, M. J. I., Neugeboren, B. I., Nelson, D. R. & Murray, A. W. Genetic drift opposes mutualism during spatial population expansion. *Proc. Natl Acad. Sci. USA* **111**, 1037–1042 (2014).
- Cao, Y. et al. Collective space-sensing coordinates pattern scaling in engineered bacteria. *Cell* **165**, 620–630 (2016).
- Cremer, J. et al. Chemotaxis as navigation strategy to boost range expansion. *Nature* <https://doi.org/10.1038/s41586-019-1733-y> (2019).
- Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
- Bosshard, L. et al. Accumulation of deleterious mutations during bacterial range expansions. *Genetics* **207**, 669–684 (2017).
- Wolfe, A. J. & Berg, H. C. Migration of bacteria in semisolid agar. *Proc. Natl Acad. Sci. USA* **86**, 6973–6977 (1989).
- Deforet, M., Carmona-Fontaine, C., Korolev, K. S. & Xavier, J. B. Evolution at the edge of expanding populations. *Am. Nat.* **194**, 291–305 (2019).
- Lenski, R. in *Microbial Ecology: Principles, Applications and Methods* (eds Levin, M. et al.) 183–198 (McGraw-Hill, 1992).
- Smith, J. M. *Evolution and the Theory of Games* (Cambridge Univ. Press, 1982).
- Levin, B. R. Frequency-dependent selection in bacterial populations. *Phil. Trans. R. Soc. Lond. B* **319**, 459–472 (1988).
- Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
- Hansen, C. H., Endres, R. G. & Wingreen, N. S. Chemotaxis in *Escherichia coli*: a molecular model for robust precise adaptation. *PLOS Comput. Biol.* **4**, e1 (2008).
- Korobkova, E., Emonet, T., Vilar, J. M., Shimizu, T. S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **428**, 574–578 (2004).
- Park, H., Guet, C. C., Emonet, T. & Cluzel, P. Fine-tuning of chemotactic response in *E. coli* determined by high-throughput capillary assay. *Curr. Microbiol.* **62**, 764–769 (2011).
- Si, G., Wu, T., Ouyang, Q. & Tu, Y. Pathway-based mean-field model for *Escherichia coli* chemotaxis. *Phys. Rev. Lett.* **109**, 048101 (2012).
- Tu, Y. Quantitative modeling of bacterial chemotaxis: signal amplification and accurate adaptation. *Annu. Rev. Biophys.* **42**, 337–359 (2013).
- Liu, C. et al. Sequential establishment of stripe patterns in an expanding cell population. *Science* **334**, 238–241 (2011).
- Merrell, D. *The Adaptive Seascape: The Mechanism of Evolution* (Univ. Minnesota Press, 1994).
- Mustonen, V. & Lässig, M. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.* **25**, 111–119 (2009).
- Poelwijk, F. J., de Vos, M. G. & Tans, S. J. Tradeoffs and optimality in the evolution of gene regulation. *Cell* **146**, 462–470 (2011).
- Towbin, B. D. et al. Optimality and sub-optimality in a bacterial growth law. *Nat. Commun.* **8**, 14123 (2017).
- Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, 1998).
- Gore, J., Youk, H. & van Oudenaarden, A. Snowdrift game dynamics and facultative cheating in yeast. *Nature* **459**, 253–256 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Media and growth conditions

The TB medium contains 10 g tryptone and 5 g NaCl per litre. The M9 supplemented medium is based on the Knight laboratory's recipe: 1× M9 salts, 0.2% casamino acids, 2 mM MgSO₄, 0.1 mM CaCl₂, and the carbon source is 0.4% (v/v) glycerol. M9 salts were prepared to be 5× M9 salts stock solution (in 1 l): Na₂HPO₄ 30 g, KH₂PO₄ 15 g, NH₄Cl 5.0 g, NaCl 2.5 g. The Luria–Bertani (LB) medium used in this study contains 2.5 g yeast extract, 5 g bacto tryptone, and 5 g NaCl per litre. For all the expansion experiments, the bacto-agar (BD, 214010) was added to the growth medium and the agar concentration varied from 0.2% to 0.3% (w/v). To prepare semi-solid agar, the above growth medium was buffered to pH 8.0 with 0.1 M HEPES (pH 8.0), the pH variation was less than 0.3. Then, 10 ml of the above medium supplemented with different agar concentrations was poured into a 90-mm Petri dish, and allowed to harden at room temperature for 90 min. Unless otherwise stated, all other reagents were from Sigma. All experiments were carried out at 37 °C. Plasmids were maintained with 100 µg/ml ampicillin, 50 µg/ml kanamycin, 25 µg/ml chloramphenicol, or 50 µg/ml spectinomycin.

Strains and plasmid construction

The ancestor *E. coli* CLM strain used in this study was provided by A. Danchin (AMAbiotics, France). All strains used in this study are listed in Supplementary Table 2. Strains are available upon request. Oligonucleotides used are listed in Supplementary Table 4.

The *cheZ*-titratable strain WL1 (*AcheZ*, *Δlac*, *bla:P_{tet}-tetR-cheZ* at *attB* site) was constructed as described previously³⁵. In brief, the *bla:P_{tet}-tetR-cheZ* feedback loop was amplified from the pMD19-T₀-Amp-T₁-P_{tet}-*tetR-cheZ* with primers PR29 and PR30 and inserted into the CL1 (*ΔcheZ*, *Δlac*) chromosomal *attB* site by recombineering with the aid of pSIM5³⁶. To construct the *cheZ* titratable strain WL2 (*AcheZ*, *Δlac*, *bla:P_{lac}-lacI-cheZ* at *attB* site) the plasmid pMD19-T₀-Amp-T₁-P_{lac}-*lacI-cheZ* was first constructed. In brief, the plasmid was constructed by inserting PCR-amplified *P_{lac}-lacI* with XhoI and HindIII restriction sites from QL5 into the corresponding restriction sites of the pMD19-T₀-Amp-T₁-P_{tet}-*tetR-cheZ*, finally the *P_{tet}-tetR* was replaced with the *P_{lac}-lacI*. Then the *bla:P_{lac}-lacI-cheZ* feedback loop was amplified from the plasmid and inserted into the *attB* site by recombineering.

The derived *clpX* (G371S, *425Q, P67S), *rcsD* (G656V), *cheB* (D37E), *mutH* (R136C) alleles were moved into the ancestor strain using the two-plasmid-based CRISPR–Cas9 system³⁷ with minor modifications. In brief, the targeted locus was first replaced with *bla* gene by λ-Red, the evolved alleles were then introduced by CRISPR–Cas9 targeting the *bla* gene. Procedures used for the allelic exchange are as follows. First, the pCASSac plasmid containing the *cas9*, λ-Red and *sacB* was electroporated into the ancestor and strains with the kanamycin resistance were selected. Then, PCR products containing the -50-bp sequence homologous to the targeted locus on each side of the T₀-P_{EM}-7-*bla*-T₁ cassette were amplified from pMD19-T₀-*bla*-T₁-P_{tet}-*tetR* plasmid. The purified DNA fragments were then electroporated into the ancestor strain with pCASSac. The T₀-P_{EM}-7-*bla*-T₁ cassette was integrated into the targeted locus with the aid of the λ-Red from pCASSac and the ampicillin-resistant clones were selected. These cells were identified by colony PCR. Subsequently, the pTargetF-AmpR plasmid carrying the N20 (20-bp region complementary to the target region) that was targeted to the *bla* region was obtained by inverse PCR with the primers PCC35 and PCC36 from the pTargetF followed by self-ligation. PCR products containing the evolved alleles were amplified from the evolved strains with appropriated primers listed in Supplementary Table 4. The pTargetF-AmpR plasmid carrying the sgRNA, *lacIq*-P_{trc}

promoter guiding the PMB1 replication of pTarget and the PCR fragments with the evolved allele were co-electroporated into the ancestor with T₀-P_{EM}-7-*bla*-T₁ cassette at the targeted locus and pCASSac plasmid. Cells grown on LB agar containing kanamycin and spectinomycin were identified by colony PCR and DNA sequencing. Finally, the constructed strains both contained the pTarget-AmpR and the pCASSac plasmids. The pTarget-AmpR was first cured by a second round of genome editing. Cells grown on LB agar with kanamycin that did not grow on LB agar with spectinomycin were selected. These selected cells were picked into 2 ml of LB medium with 5g/l glucose for overnight culture; cells grown on LB agar containing 5g/l glucose and 10g/l sucrose were selected as the final strains.

The fluorescence plasmid PZA31-Ptet-M2-GFP was from the Hwa laboratory. To construct a loss-of-function non-fluorescent GFP mutant NFP³⁸, the 66th amino acid Y was mutated into C by overlapping PCR. In brief, PZA31-Ptet-M2-GFP was reverse amplified with a pair of complementary primers GFP-Y66C-f and GFP-Y66C-r. The PCR product was purified and treated with DpnI, gel purified, ligated, and then transformed into the DH5α-competent cells. The plasmid was verified by sequencing.

Evolution experiment procedures

First, the ancestor strain from the -80 °C stock was streaked onto an agar plate and cultured at 37 °C overnight. Three to five single colonies were picked into 2-ml of the corresponding growth medium and cultured at 37 °C overnight. Second, the overnight culture was diluted into 2-ml pre-warmed fresh growth medium with a ratio of 1:100 the next morning. Bacteria were then cultured to the mid-log phase (OD₆₀₀ was around 0.2–0.3), and 2 µl of the ancestor strain was inoculated at the centre of a semi-solid agar plate and incubated at 37 °C for 24 h. Cells grew and migrated, occupying the whole semi-solid agar plate (marked as cycle 0). We picked 2 µl of the agar–cell mixture from site A, B, C, D, or E of this master plate (with a radius of 5, 10, 15, 20, or 25 mm away from the inoculum, respectively), directly inoculated onto fresh semi-solid agar plates (marked as A, B, C, D, or E series, correspondingly), and incubated at 37 °C for another 24 h, marked as cycle 1. Then 2 µl of the agar–cell mixture was picked at site A from plate A, site B from plate B, site C from plate C, site D from plate D, and site E from plate E, inoculated onto the centre of the fresh semi-solid agar plates, and incubated at 37 °C for 24 h, marked as cycle 2. This process was repeated for 50 cycles, with selection always kept at the same radius. Bacteria from site A', A, B, C, D, or F (with a radius of 3, 5, 10, 15, 20, or 35 mm away from the inoculum, respectively) were passaged for 40–50 cycles following the above process in M9 minimal medium supplemented with glycerol and casamino acid medium (M9 + glycerol + CAA). For the evolution experiment carried out in semi-solid LB agar plates, 2 µl of cell–agar mixture picked from site A, D, or F (with a radius of 5, 20, or 35 mm away from the centre of the semi-solid agar plate, respectively) were transferred onto fresh semi-solid agar plates. Considering the fast growth rate and expansion speed in this medium, the cycle of the culture was shortened to 12 h. For the evolution experiment carried out in M9 minimal medium supplemented with glycerol, A', D, or F (with the radius of 3, 20, or 35 mm away from the centre of the semi-solid agar plate, respectively) were transferred every 72 h, and this process was repeated for 30 cycles. The samples of the evolving populations were stored at -80 °C right after well mixing the agar–cell mixture from indicated sites with an equal volume of 40% (v/v) glycerol. All the evolution experiments were performed in at least three replicates. The 5 × 10³-fold daily growth corresponds to -12.3(log₂[5 × 10³]) generations of doubling. The number of doublings during each cycle was estimated as follows: cell density was about 3.54 × 10⁹/ml in TB medium and 1.93 × 10⁹/ml in M9 + glycerol + CAA after migrating in semi-solid agar for 24 h. The initial inoculum cell density was about 7.43 × 10⁵ per ml for TB and 4.94 × 10⁵ per ml for M9 + glycerol + CAA. These numbers were counted using flow cytometry (see below).

Expansion speed measurement

The semi-solid agar plate was illuminated from below by a circular white LED array with a light box as described previously²⁸ imaged at 30 min or 1 h intervals using a Canon EOS 600D digital camera. Images were analysed using ImageJ and a custom-written image analysis script using MATLAB. A circle was fitted to the intensity maximum in each image and the area (A) of the fitted circle was determined. The radius (r) of the colony was calculated as $r = \sqrt{A/\pi}$. The maximum expansion speed was calculated using a linear fit over a sliding window of at least four time points, with the requirement that the fit has an r^2 greater than 0.99.

Growth rate measurement

Growth rates of the evolved strains were measured in a 100-ml flask with 20 ml corresponding growth medium at 37 °C, 150 rpm. The procedure was as follows. First, the isolated bacteria from –80 °C stock was streaked onto the agar plate, and cultured at 37 °C overnight. Second, 3–5 single colonies were picked and inoculated into 2 ml corresponding growth medium and cultured overnight. The overnight culture was diluted into 2 ml pre-warmed medium with a ratio of 1:100 the next morning and cultured to log phase. The log phase culture was successively diluted into 20 ml pre-warmed growth medium, the final OD₆₀₀ was about 0.02–0.05. OD₆₀₀ was measured using a spectrophotometer reader every 12 min (for TB) or 15 min (for M9). At least three doubling times were recorded. Maximum growth rates were calculated using an exponential fit over a sliding window of at least five time points, with the requirement that the fit has an r^2 greater than 0.99.

Competition assay

Head-to-head chemotactic competition was observed using a pair of strains with competition either between the ancestor strain and one evolved strain, or between the two *cheZ*-titratable strains (Extended Data Fig. 3) induced with aTc or IPTG. To allow observation by fluorescence microscopy, strains carrying either GFP- or NFP-expressing plasmids were prepared (Extended Data Fig. 3). Each competition was repeated for both combinations of plasmids (for example in the competition run ancestor versus A, the growth and expansion of the ancestor strain was observed for the run with Anc_G versus A_N as well as for the run Anc_N versus A_G). Both the fluorescence intensity and the cell number of each pair of competing strains at different positions and different time points were measured to characterize the competitiveness of the cells in semi-solid agar. The competition experiments for ancestor versus an evolved strain were initiated as follows: three to five single colonies of the isolated evolved strain and the ancestor with the GFP/NFP were cultured to log phase (OD₆₀₀ was around 0.20) separately. Two types of the combined mixed strains were prepared: the evolved strain with GFP was mixed with the Anc_N while the evolved strain with NFP was mixed with the Anc_G in a 1:1 ratio. Next, 2 µl of the two types of the combined mixture were inoculated onto the centre of pre-prepared semi-solid agar plates separately and allowed to expand at 37 °C. The fluorescence intensity of the evolved strains or ancestor with the GFP reporter from these two plates after the expansion were scanned by a Nikon Ti-E microscope equipped with a 10× phase-contrast objective (NA = 0.30) and a Andor Zyla 4.2 s CMOS camera. The fluorescence intensity was used to represent the bacterial density in semi-solid agar. Samples were also collected before and after the expansion. Images were collected from the Z axis position at the fixed plane where the bacteria's fluorescence was constant in the semi-solid agar. Each image was collected at a height of 1 mm horizontally across the plate and 100 images were collected. The fluorescence intensity was calculated using NIS-Elements AR 4.50 software. The competition between different *cheZ*-titratable strains was performed following the same protocol with different concentrations of the aTc and IPTG added to the growth medium and semi-solid agar.

Fluorescence intensity as a function of cell density was calibrated as follows. The Anc_G, A_G, B_G, C_G, D_G and E_G strains were cultured in TB medium to mid-log phase; OD₆₀₀ was around 0.20 for each strain. A total of 200 ml culture was collected and concentrated to 1.6×10^{10} cells/ml in the TB medium with 2 mg/ml kanamycin for each strain. Then, a serial dilution of the above concentrated sample into TB medium with 2 mg/ml kanamycin was carried out. Subsequently, the diluted samples were mixed with 0.277% (w/v) TB agar containing 2 mg/ml kanamycin in a ratio of 1:9; 10 ml of the cell–agar mixture was poured into a 9-cm Petri dish and allowed to solidify at room temperature for 90 min, and 100 µl of the cell–agar mixture was used for cell counting with a flow cytometer. The fluorescence intensity of the above cell–agar mixture plate was measured using a fluorescence microscope, as for the two-strain competition assay. Then the relationship between the fluorescence intensity of the cell in semi-solid agar and the cell density was plotted.

The initial and final ratios of the two competitors were measured by cell counting with a flow cytometer (Beckman, Cyto-FLEX). In brief, samples were first fixed with pre-cooled cell counting buffer (0.9% NaCl with 0.12% formaldehyde). Subsequently, the fixed samples were diluted as necessary with straining buffer (cell counting buffer with 0.1 µg/ml DAPI) before the flow cytometer analysis. Finally, the strained samples were counted with the flow cytometer. The flow rate was 30 µl/min and at least 50,000 cells were collected. The DAPI-stained particles were deemed to be the bacterial cells, and the DAPI-positive cells were separated into two groups (GFP and NFP) through the FITC channel. The fitness W_i of strain i (relative to the ancestor) is defined as the ratio of density at distance d (and a sufficiently long time t) over the initial inoculant density, $\rho_x(d,t)/\rho_x(0,0)$, relative to the same ratio for the ancestor: $W_i(d) = [\rho_x(d,t)/\rho_x(0,0)]/[\rho_{anc}(d,t)/\rho_{anc}(0,0)]$.

Quantitative real-time RT-PCR

A volume of 1 ml of the log phase bacteria (OD₆₀₀ ~0.2) from each condition was immediately mixed with 2 ml RNA protect Bacteria Reagent (Qiagen). Total RNA was extracted using the RNeasy Mini kit (Qiagen) according to the manufacturer's protocol. The RNA yield and purity were checked using a NanoDrop 2000c spectrophotometer (Thermo Scientific), and the absence of genomic DNA contamination was confirmed by PCR. About 500 ng RNA was reverse transcribed, using a PrimeScript RT reagent kit with gDNA Eraser (Takara) according to the manufacturer's protocol. Reactions without reverse transcriptase were conducted as controls for the following qPCR reactions. Then the cDNA samples were diluted 1:25 with PCR-grade water and stored at –20 °C until use. SYBR Premix Ex Taq (Tli RNaseH plus) (Takara) was used for qPCR amplification of the cDNA. Then, 5 µl diluted cDNA sample, 200 nM forward and reverse qPCR primers, 10 µl SYBR Premix Ex Taq, and up to 20 µl PCR-grade water were mixed in a well of a Hard-Shell 96-well PCR plate (BIO-RAD). The non-template control (NTC), containing sterile water instead of cDNA template, was included during each qPCR experiment to check the purity of the reagents. Each reaction was performed in triplicate. The qPCR reactions were performed using a Bio-rad CFX connect Real-Time system with the following program: 30 s at 95 °C and 40 cycles of denaturation (5 s at 95 °C), annealing, and elongation (30 s at 60 °C). Data were acquired at the end of the elongation step. A melting curve was run at the end of the 40 cycles to check the specificities of the accumulated products. To calculate PCR efficiency, standard curves were made for the target gene by using serially diluted cDNA samples as the templates. 16S rRNA was used as the reference gene to normalize the expression level.

Single-cell tracking

Single-cell tracking was performed as described previously³⁹ with minor modifications. A custom MATLAB script was used to control the automated stage of the microscope (Nikon Ti-E) via the MicroManager interface⁴⁰. Movies were sequentially acquired using an Andor Zyla 4.2 sCMOS camera at 10 frames/s through a 10× phase-contrast

objective (NA = 0.30), 1,024 pixels by 1,024 pixels for 1 min. All movies were analysed using a custom-written MATLAB code as described previously^{39,41}. The single-cell tracking was carried out as follows: first, three to five single colonies of the isolated evolved strain or 5 µl of the evolved cell-agar mixture stored at -80 °C were cultured in 2 ml growth medium at 37 °C, 150 rpm overnight. Second, the overnight culture was diluted into 2 ml pre-warmed growth medium at a 1:100 ratio, cultured at 37 °C until the OD₆₀₀ reached 0.2–0.3. This step was then repeated, and cells were diluted into 15 ml pre-warmed growth medium. Bacteria were harvested until the OD₆₀₀ reached 0.2. Third, the harvested sample was diluted with pre-warmed growth medium to a final cellular density of OD₆₀₀ = 0.05. Then, 5 µl of the diluted sample was pipetted onto a microscope glass slide (25 mm × 75 mm), and a coverslip (18 mm × 18 mm) was placed onto the top of the sample (slowly and with caution to avoid the formation of air bubbles). Subsequently, sides of the coverslip were sealed with heated wax in order to avoid evaporation. The growth medium used for single-cell tracking was supplemented with 0.05% (w/v) PVP 40,000 to protect the flagella. Two to three slides were prepared for each strain, and five 1-min-long videos were acquired. Temperature was kept at 37 °C during video acquisition.

Whole-genome sequencing and analysis

Whole-genome sequencing was performed using the Illumina platform, obtaining an average >100× coverage. Samples directly collected from the TB evolution experiment in 0.25% agar were used for sequencing. In brief, the isolated clones were grown in the growth medium as described previously and harvested at stationary phase. For the population samples, 100 µl of the frozen cell-agar mixture from the 50th cycle was grown in 3 ml growth medium and harvested at the 10th hour. Cellular DNA was extracted using a genomic DNA purification kit (Tiangen) according to the manufacturer's protocol. Whole-genome libraries were prepared and sequenced on the Illumina HiSeq X10 by BGI. The genome sequence of *E. coli* str. K-12 substr. MG1655 (NCBI: NC_000913.3) was used as the reference sequence. All sequencing data were analysed using the BRESEQ pipeline⁴² supported by Python. A subset of the identified mutations was resequenced using Sanger sequencing for confirmation.

Numerical simulations of competition dynamics and evolution

The growth–expansion model¹⁴ (Extended Data Fig. 5) was extended to analyse competition and evolutionary dynamics during expansion, see Extended Data Fig. 6 for introduction and defining partial differential equations (PDE), and Supplementary Model for mathematical details. Numerical solution of the partial differential equations was done using an implicit scheme using Python 2.7 and the PDE solver module FiPy⁴³. Integration over time was typically performed with time steps $dt = 0.25$ s, and a grid resolution with spacing $dx = 10$ µm. Simulations were performed using a custom-made Python code, which is available via GitHub at https://github.com/jonascremer/chemotaxis_simulation. Used parameter sets are provided in Supplementary File simulationparameters.txt, see Supplementary Model 5 for usage.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Sequencing data have been deposited to the NCBI Sequence Read Archive (SRA), accession PRJNA559221. Other major experimental data supporting the findings of this study are available within the paper and Supplementary Information. Simulation data can be generated with the custom-made code and the parameter sets provided.

Code availability

Custom-made simulation code is available via GitHub at https://github.com/jonascremer/chemotaxis_simulation.

35. Zheng, H. et al. Interrogating the *Escherichia coli* cell cycle by cell dimension perturbations. *Proc. Natl Acad. Sci. USA* **113**, 15000–15005 (2016).
36. Datta, S., Costantino, N. & Court, D. L. A set of recombinering plasmids for Gram-negative bacteria. *Gene* **379**, 109–115 (2006).
37. Jiang, Y. et al. Multigene editing in the *Escherichia coli* genome via the CRISPR–Cas9 system. *Appl. Environ. Microbiol.* **81**, 2506–2514 (2015).
38. Fu, J. L., Kanno, T., Liang, S.-C., Matzke, A. J. M. & Matzke, M. GFP loss-of-function mutations in *Arabidopsis thaliana*. G3 **5**, 1849–1855 (2015).
39. Waite, A. J. et al. Non-genetic diversity modulates population performance. *Mol. Syst. Biol.* **12**, 895 (2016).
40. Edelstein, A. D. et al. Advanced methods of microscope control using µManager software. *J. Biol. Methods* **1**, e10 (2014).
41. Dufour, Y. S., Gillet, S., Frankel, N. W., Weibel, D. B. & Emonet, T. Direct correlation between motile behavior and protein abundance in single cells. *PLOS Comput. Biol.* **12**, e1005041 (2016).
42. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* **1151**, 165–188 (2014).
43. Guyer, J. E., Wheeler, D. & Warren, J. A. FiPy: partial differential equations with Python. *Comput. Sci. Eng.* **11**, 6–15 (2009).
44. Fisher, R. The wave of advance of advantageous genes. *Ann. Eugen.* **7**, 355–369 (1937).
45. Kolmogorov, A. N., Petrovsky, I. & Piscounov, N. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Mosk. Univ. Bull. Math.* **1**, 37 (1937).
46. Korolev, K. S. Evolution arrests invasions of cooperative populations. *Phys. Rev. Lett.* **115**, 208104 (2015).
47. Yang, F., Moss, L. G. & Phillips, G. N. Jr. The molecular structure of green fluorescent protein. *Nat. Biotechnol.* **14**, 1246–1251 (1996).
48. Keller, E. F. & Segel, L. A. Model for chemotaxis. *J. Theor. Biol.* **30**, 225–234 (1971).
49. Vaknin, A. & Berg, H. C. Physical responses of bacterial chemoreceptors. *J. Mol. Biol.* **366**, 1416–1423 (2007).
50. Taylor, J. R. & Stocker, R. Trade-offs of chemotactic foraging in turbulent water. *Science* **338**, 675–679 (2012).
51. Fu, X. et al. Spatial self-organization resolves conflicts between individuality and collective migration. *Nat. Commun.* **9**, 2177 (2018).

Acknowledgements We thank L. Chao, X. Fu, X. He, J.-D. Huang, A. Murray, M. Vergassola, C.-I. Wu and G. Zhao for discussions, and Y. Wu and H. Zhou for assistance with bioinformatic analyses. C.L., W.L. and D.L. acknowledge financial support by the Major Research Plan of the National Natural Science Foundation of China (91731302), National Key Research and Development Program of China (2018YFA0902700), Strategic Priority Research Program (XDB29050501), Key Research Program (KFZD-SW-216) of Chinese Academy of Sciences, and Shenzhen Grants (JCYJ20170818164139781, KQTD2015033117210153, Engineering Laboratory [2016]1194). T.H. and J.C. acknowledge support from the NIH through grant R01GM95903.

Author contributions C.L. initiated and directed the research. W.L. and D.L. carried out most of the experiments, with contributions from C.L. and J.C. J.C. and T.H. developed the model and carried out the numerical simulations and mathematical analysis. All authors analysed the results. J.C., T.H. and C.L. wrote the manuscript with contributions from W.L.

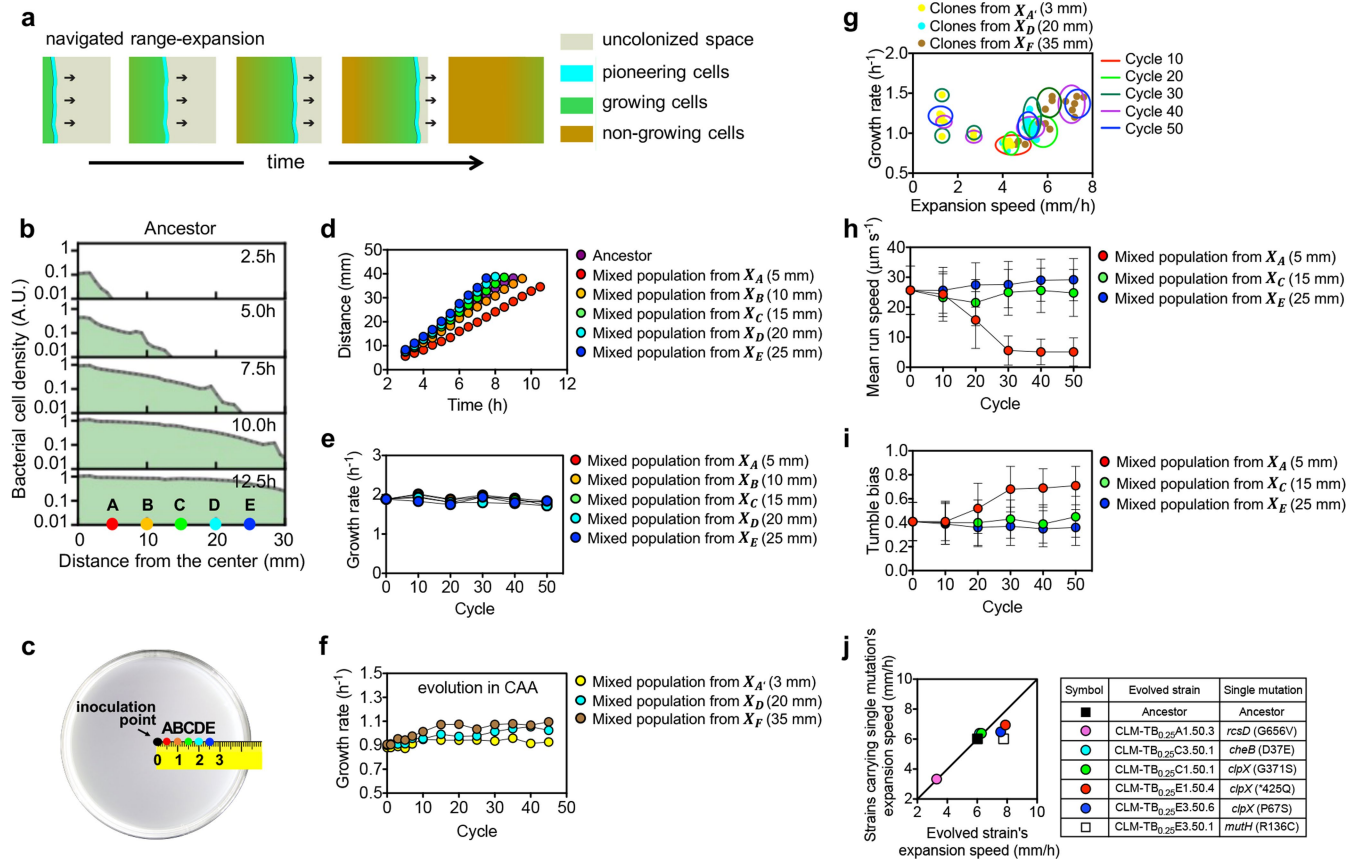
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1734-x>.

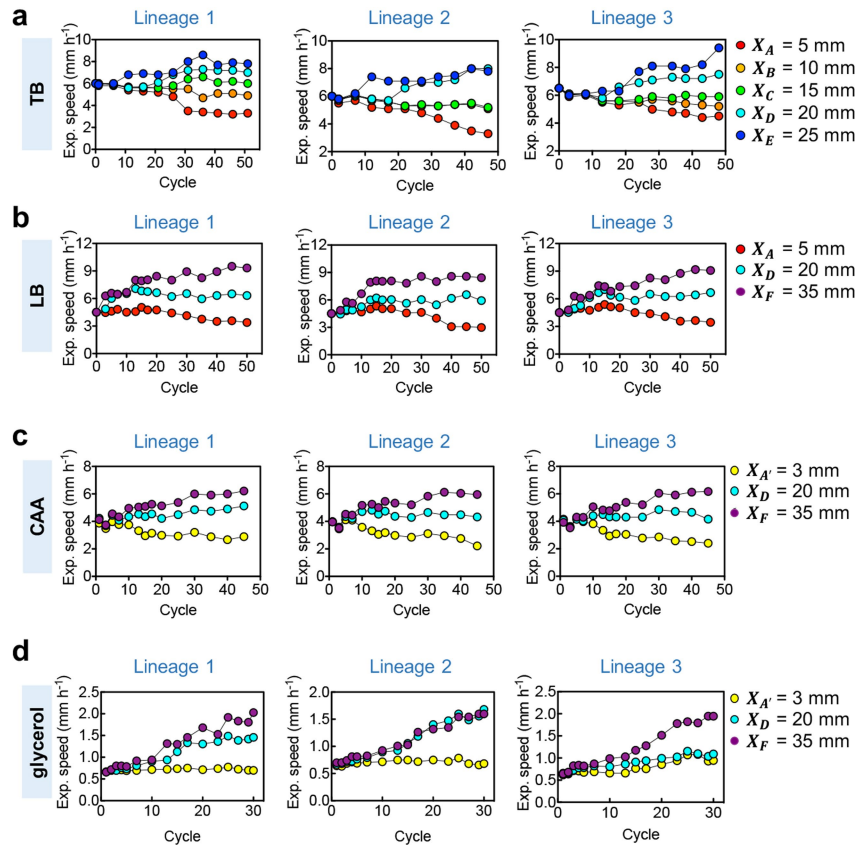
Correspondence and requests for materials should be addressed to T.H. or C.L.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



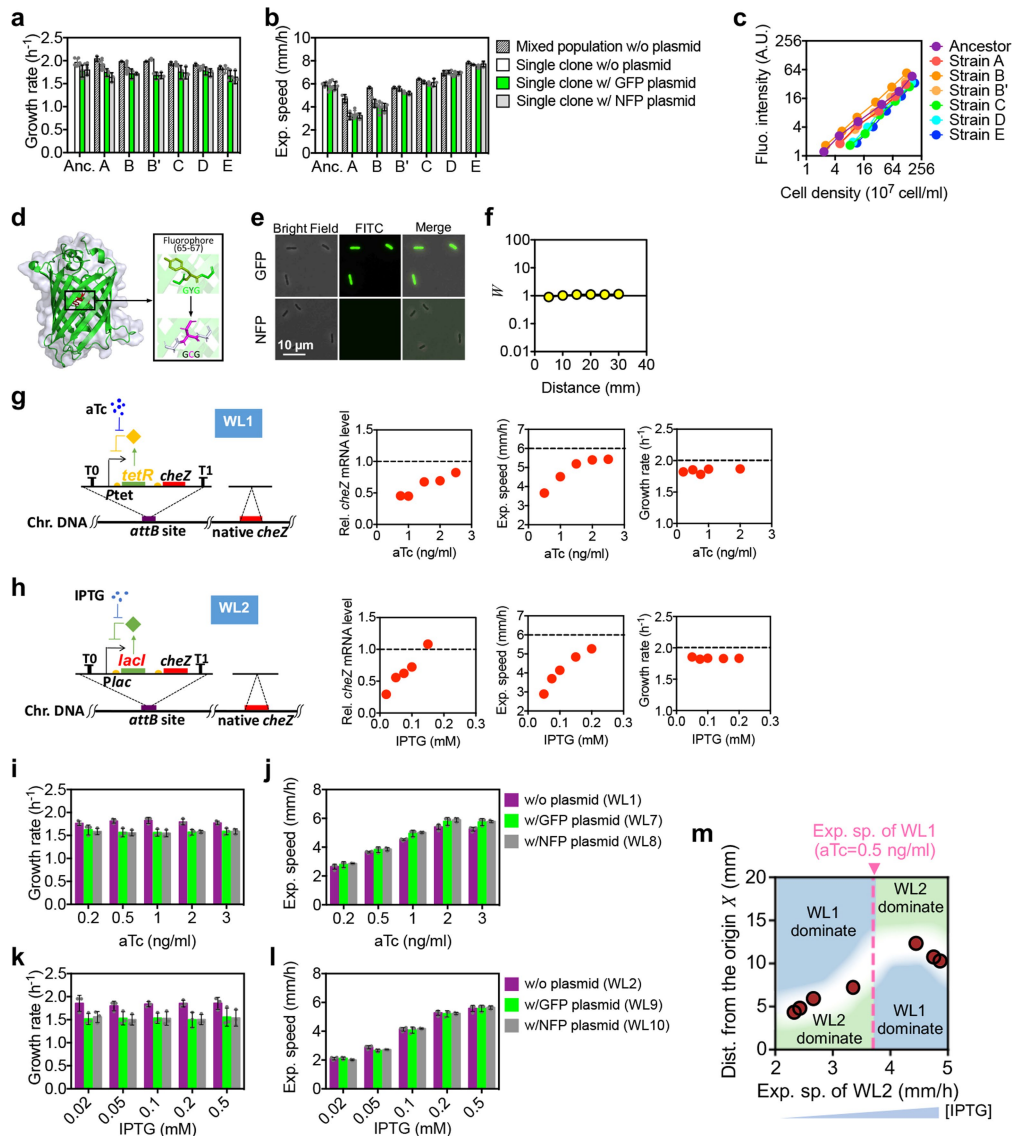
Extended Data Fig. 1 | Characteristics of cell motility and expansion speed for the ancestor and for mixed populations of evolved cells. **a**, Navigated range expansion by bacteria¹⁴ on a soft agar plate involves a group of pioneering cells (cyan) at the population front that move outwards towards uncolonized space (grey) in response to a signalling cue. During the outward migration, they replicate and leave offspring behind the front. The offspring do not move outwards, but settle wherever they are deposited and grow exponentially (green) until they reach carrying capacity (brown). **b**, Density profile of a population of ancestor cells containing GFP (Anc_c) at various times after inoculation at the centre of a semi-solid TB plate with 0.25% agar incubated at 37 °C. The population expanded at a defined expansion speed, covering positions A–E well before selection at 24 h. The density profiles were obtained using confocal microscopy as described previously¹⁴. Cellular characteristics were not significantly affected by the expression of GFP (Extended Data Fig. 3). **c**, The experiment described in Fig. 1a was carried out independently for five distinct selection positions, at distances X_A , X_B , X_C , X_D and X_E , ranging from 5 to 25 mm from the centre. Three independent lineages of this experiment were propagated in parallel (Extended Data Fig. 2a). Selected samples at cycle n of series S in lineage l are referred to as SLn in Supplementary Table 2. **d**, Front position versus time for the ancestral strain CLM (purple) and populations of evolved cells obtained at the 50th cycle of each of the five evolution series from lineage 1. For each mixed population, collected cells were grown to mid-exponential phase ($OD_{600} = 0.20$), and 2 μ l of the batch culture

was inoculated at the centre of the same TB plate as in **b** and incubated in the same way. The agar plates were photographed at different times after inoculation and the radius of each expanding population was deduced from the area measured using ImageJ. The expansion speed of each population (Fig. 1b) was obtained as the slope of the linear fit of the data after $t = 3$ h. **e**, Growth rates of population samples from each of the five selection series from lineage 1 at various evolution cycles. **f**, Growth rates of population samples from each of the three selection series (lineage 1) at various cycles of evolution experiments in CAA medium. **g**, Scatter plot of growth rates and expansion speeds for single strains isolated from frozen samples taken at various cycles of CAA evolution experiments. The evolution cycles are indicated by circles. **h**, **i**, Mean run speed (**h**) and tumble bias (**i**) of population samples from each of the three selection series from lineage 1 of evolution experiments carried out in TB medium at various evolution cycles. At least 10,000 cells were subjected to single-cell tracking analysis for each experiment (see Methods). Error bars represent s.d. **j**, Expansion speeds of several strains, each carrying an identified mutation, plotted against the corresponding evolved strains from which the mutations were identified. The mutations are indicated in the legend (see Supplementary Table 1). *mutH* does not affect motility and serves as a control. Experiments were repeated independently three times for **b**, **d** and twice for **e**, **f** with similar results. In **h**, **i**, data are mean \pm s.d. for a single biological replicate, $n = 10,000$ cells analysed. In **j**, data are means for $n = 3$ biological replicates (s.d. error bars are smaller than the symbols).



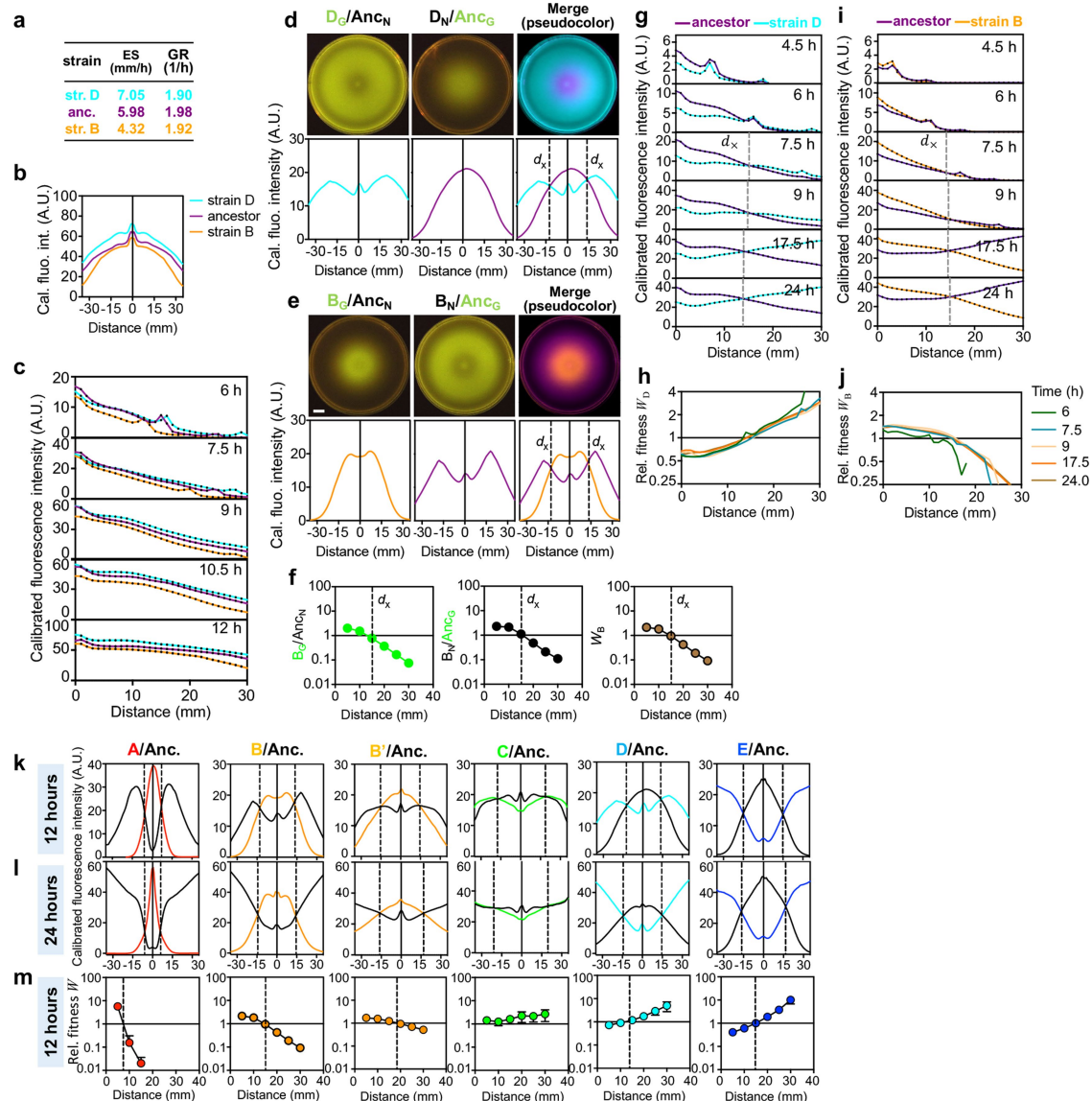
Extended Data Fig. 2 | Experimental evolution of expansion speed in different growth media. **a–d**, Expansion speeds of population samples from each selection series are shown at various cycles of evolution experiments carried out in TB (**a**), LB (**b**), CAA (**c**) and glycerol-containing minimal medium (**d**). The agar density was 0.25% (w/v). All three lineages of each medium showed similar trajectories. The absence of a chemoattractant in the glycerol minimal medium leads to a very different outcome compared to the experiments carried out in complex media (**a–c**). In the absence of a chemoattractant, cells

follow simple Fisher–Kolmogorov dynamics^{14,44,45}. In line with the absence of a growing population trailing the front, no decrease in swimming behaviour was observed over time. However, as previously investigated, slower swimming behaviour might be selected for when density-dependent growth effects or strong tradeoffs between swimming and cell growth exist^{18,46}. Experiments in **a, b, d** were carried out once with three biologically independent repeats with similar results; experiment in **b** was repeated independently twice with similar results.



Extended Data Fig. 3 | Fitness effects for cells expressing GFP and its non-fluorescent variant NFP. **a, b**, Growth rates (**a**) and expansion speeds (**b**) of the ancestor and six mutant clones harbouring GFP, NFP, or no plasmid, along with the corresponding mixed populations from which the six mutant clones were isolated. Plasmids harbouring constitutive GFP or NFP expression (PZA31-Ptet-M2-GFP or PZA31-Ptet-M2-NFP, respectively) were transformed into the six evolved strains A, B, B', C, D, and E, to form A_G , A_N and so on (Supplementary Table 2). **c**, Fluorescence intensity as a function of cell density. The cell growth of cells expressing GFP was arrested by adding 2 mg ml^{-1} kanamycin, and cells were concentrated to 1.6×10^{10} cells per ml. Subsequently, serial dilutions were carried out. For each cell density, cells were vigorously mixed with pre-warmed 0.277% (w/v) TB agar containing 2 mg ml^{-1} kanamycin at a ratio of 1:9 and poured into three Petri dishes with 10 ml each. All dishes were allowed to harden at room temperature for 90 min. The cell–agar mixture was subjected to cell counting by fluorescence-activated cell sorting (FACS) and the fluorescence intensity of the agar plate was detected using a fluorescence microscope. **d**, The three-dimensional structure of GFP with the predicted position of the loss-of-function mutation introduced, Y66C. The shown crystal structure is based on the NFP protein sequence aligned to PDB ID: 1GFL⁴⁷. The rectangle shows the mutation position. **e**, Ancestor cells harbouring constitutive GFP or NFP expression (A_{G_0} and A_{N_0} cells, respectively) as viewed by fluorescence microscopy; the images verify the loss-of-function mutation of GFP. **f**, The relative fitness W of A_{G_0} and A_{N_0} cells at different distances from the agar plate centre. Cells harbouring the GFP or NFP plasmid were equally mixed and

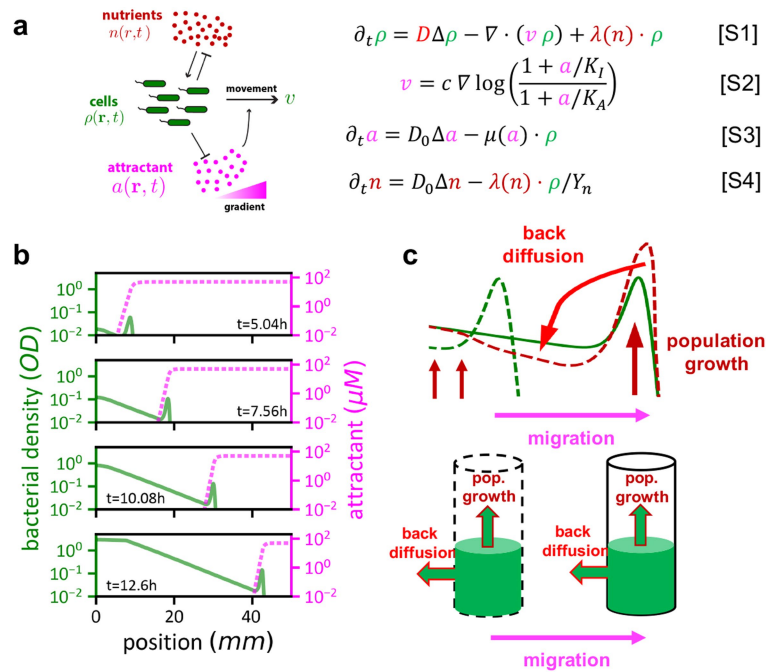
inoculated at the centre of 0.25% TB plates. Cell–agar mixtures picked at various distances were subjected to cell counting by FACS 24 h after inoculation and the ratio was reported as W (see Methods). **g, h**, The genetic circuit and characteristics of the *cheZ*-titratable strains WL1 (**g**) and WL2 (**h**). The expression of *cheZ* is under the control of $P_{\text{tet}}\text{-tetR}$ (WL1) or $P_{\text{lac}}\text{-lacI}$ (WL2) feedback loop and the native *cheZ* was seamlessly removed (see Methods). Relative *cheZ* mRNA levels change around twofold under different concentrations of inducers (aTc for WL1 and IPTG for WL2). Relative *cheZ* mRNA expression levels, expansion speed, and growth rates of WL1 (**g**) and WL2 (**h**) under various concentrations of the respective inducers are shown next to the circuits. Horizontal dashed lines show the corresponding values for the ancestor strain. **i–l**, Growth rates (**i, k**) and expansion speeds (**j, l**) of the aTc-titratable *cheZ* strain WL1 (**i, j**) and its derivatives expressing GFP (WL7) or NFP (WL8), and the IPTG-titratable *cheZ* strain WL2 (**k, l**) and its derivatives expressing GFP (WL9) or NFP (WL10). **m**, Circles indicate crossover distances between fluorescent derivatives of two *cheZ*-titratable strains (WL1 and WL2), with WL1 strains induced at a fixed concentration of its inducer (0.5 ng ml^{-1} aTc), and WL2 strains induced at various IPTG concentrations (expansion speeds (**h, l**) shown on the x axis). The background colours again indicate dominance by WL1 (purple) or WL2 (green). In **a, b, g–l**, data are mean \pm s.d. for $n = 3$ biologically independent repeats (individual data points shown as circles). Error bars in **g, h** were smaller than the symbols. Experiments shown in **c, e, f, m** were repeated independently three times with similar results.



Extended Data Fig. 4 | Results of two-strain competitions. **a**, Expansion speeds and growth rates of ancestor, strain B, and strain D (Supplementary Table 2).

b, Calibrated fluorescence intensity profiles of singly grown strains 12 h after inoculation at the centre of 0.25% TB agar plates. Unless noted otherwise, fluorescence intensity is normalized according to the calibration curve shown in Extended Data Fig. 3c; the relative value 1 refers to 5×10^7 cells per ml. **c**, Relative fluorescence intensities obtained at various times for the fluorescent mutant strains B_G (orange) and D_G (cyan) and the ancestor Anc_G (purple), each grown singly on TB plates. The faster strain has higher fitness everywhere. **d**, Raw photographs (top) and fluorescence intensity profiles (bottom) before and after merging of a representative two-strain competition between the fluorescent derivatives of the ancestor and strain D 12 h after initial equal inoculation. We used plasmids GFP and NFP in this study to distinguish the two strains from each other in the head-to-head competition, as there is no systematic influence caused by the expression of GFP or NFP (Extended Data Fig. 3). Top (from left to right): competition between D_G and Anc_N , competition between D_N and Anc_G , and merged photograph in pseudocolor (D_G , cyan; Anc_G , purple). Bottom: corresponding relative fluorescence intensity profiles. **e**, As in **d**, but with strain B instead of strain D. Evolved strains and the ancestor with GFP or NFP were cultured to log phase separately. Two types of the combined mixed strains were prepared and the evolved strain with GFP was mixed with Anc_N while the evolved strain with NFP was mixed with Anc_G in a 1:1 ratio. Subsequently, 2 μ l of the two types of combined mixture was inoculated onto the centre of pre-warmed semi-solid agar plates separately and allowed to expand at 37 °C for up to 24 h. Photographs and fluorescent intensities of the evolved strains and ancestor with the GFP reporter from these two plates after the expansion were taken at various

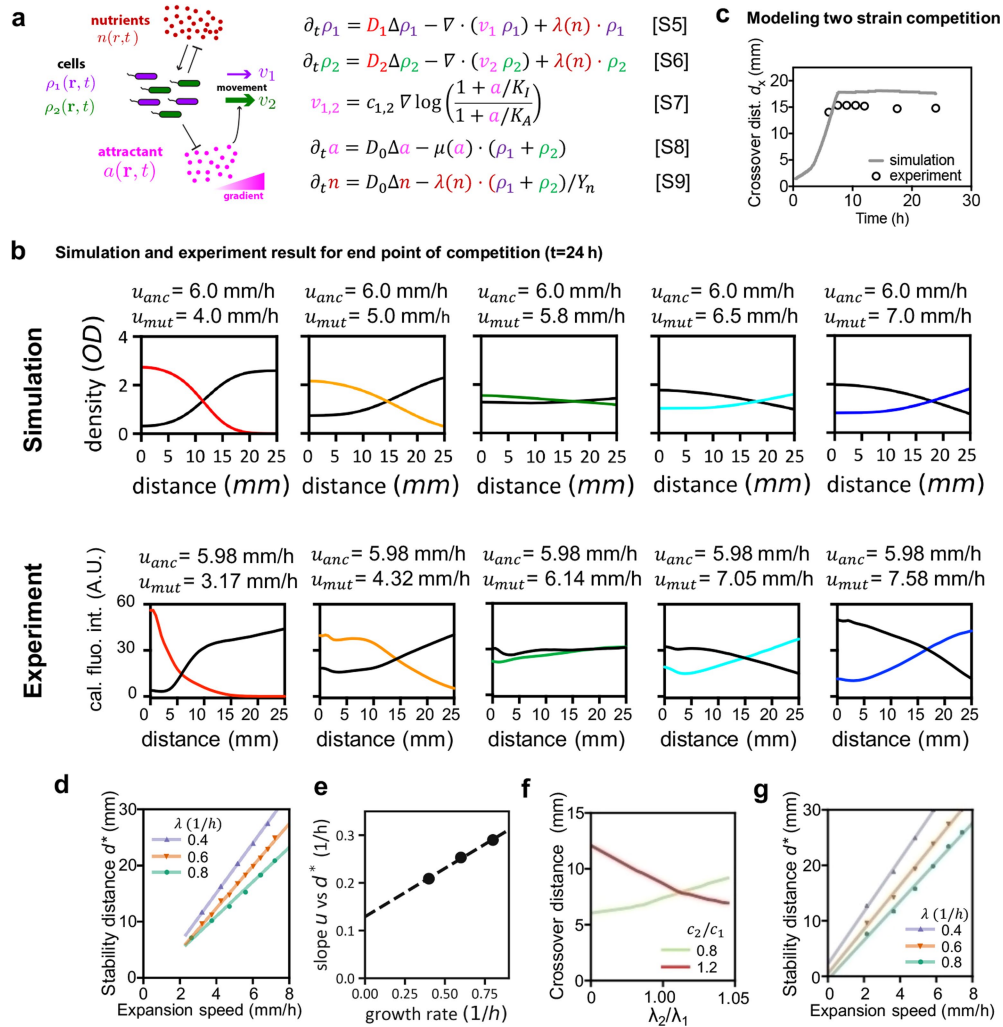
times and merged (see Methods). **f**, Relative fitness W_B obtained as a ratio of the direct cell count of the fluorescent derivatives of the ancestor and strain B, inoculated at the centre of an agar plate. From left to right: competition between B_G and Anc_N , competition between B_N and Anc_G , and averaged curve of both (the relative fitness W_B). The competition experiments are the same as in **e**. The initial and final ratios of the two competitors were measured by cell counting with a flow cytometer (see Methods). **g**, **i**, The spatiotemporal development of the bacterial density profiles, indicated by fluorescence intensities, for the competition between the ancestor (purple) and strain D (cyan; **g**) or strain B (orange; **i**) at various times during the 24-h competition. Beyond the initial period (~ 6 h), the crossover distance could be clearly defined (vertical dashed line) and was practically time-independent, with the slower strain gaining advantage in the interior and the faster strain gaining advantage in the exterior. **h**, **j**, Relative fitness values W_D (**h**) and W_B (**j**) taken as the ratio of the fluorescence intensities (mutant:ancestor) at various distances for 6 h and beyond. The vertical dashed lines indicate the crossover distance d , where the densities of the two competing strains are equal ($W_i = 1$). Thus, the crossover distance was fixed shortly after the initial period. **k–m**, Fluorescence intensity profiles (**k**, **l**), and relative fitness W (**m**) of representative two-strain competitions between the ancestor (black solid line) and evolved isolates. The data were taken 12 h (**k**, **m**) or 24 h (**l**) after co-inoculation of equal initial mixtures of the two competing strains at the centre of 0.25% TB agar plates, showing that the slower strain spatially outcompetes the faster strain within the crossover distance d , (dashed lines). See Supplementary Table 2 for strain information. Experiments in **a–l** were repeated independently three times with similar results. In **m**, the mean \pm s.d. of $n = 3$ biologically independent repeats is shown.



Extended Data Fig. 5 | The growth–expansion dynamics of a single strain.

a, We review here the GE model¹⁴ generated to describe the dynamics of a single strain of *E. coli* colonizing a soft agar plate. The GE model considers three variables and their dynamics in space x and time t : (i) cell density $\rho(x, t)$; (ii) the concentration $a(x, t)$ of an attractant that cells can sense and move towards; and (iii) the concentration $n(x, t)$ of a nutrient that cells consume to grow. Following the spirit of the classical model introduced by Keller and Segel⁴⁸ (the KS model), cells can move in a random, undirected manner (effective diffusion constant D and diffusion term in equation (S1)) and along the gradient of the signalling molecule (the attractant gradient represented by the convection term in equations (S1) and (S2)), which is generated by cellular consumption (equation (S3)). To account for observed density profiles and their evolution over time, three additional aspects beyond the KS model are important to consider. First, cells can detect and respond to attractant gradients in a scale-free manner only within a limited range between K_I and K_A , the lower and upper cutoffs describing the molecular limitations of attractant sensing^{26,49} as specified in equation (S2). D_0 denotes the molecular diffusion of the attractant and nutrient within the agar. The chemotactic parameter c denotes how cells translate the detected attractant gradient into directed movement. Second, cells grow throughout the expansion process, as described by the growth term in equation (S1), with growth rate λ . Third, growth relies not on the presence of the attractant but on the presence of nutrients. We model the latter dependence by the nutrient field $n(x, t)$ and a yield factor Y (equation (S4)). The distinct treatment of the roles of nutrient and attractant is designed to model the dynamics of bacterial cells in complex media, where non-chemotactic components in the complex media are designated as the nutrient, while the (minor) chemotactic components of the complex media are reflected by a low concentration of a single attractant; a detailed discussion and validation of the model, including comparison to other models of chemotactic migration, has been published previously¹⁴. **b**, Emerging density profile (green solid line) and

attractant concentration (magenta dashed line) in the GE model. At the front is a density bulge or peak, within which the attractant profile drops steeply, guiding cells to do chemotaxis. Directed movement of cells in the front bulge is coupled to an exponentially rising trailing region. In this region, cells grow and swim randomly, but there is no directed cell movement there owing to the low attractant concentration ($a < K_I$). Note that in cases in which multiple attractants are present in the medium, the population typically exhibits multiple rings, one corresponding to the exhaustion of each attractant. For these cases, the trailing region involving cell growth but no chemotaxis would correspond to the region inside the innermost ring, after the exhaustion of the last attractant. Thus, our model with a single attractant does not attempt to describe the movement of the outer rings, but models the transition of the density profile from the innermost ring to the exponential trailing region. As we describe in Supplementary Analyses 1, 2, the dynamics in this transition region determine strain dominance in multi-strain competition processes. **c**, Stable expansion of the population can be explained by a balance between growth of cells in the front bulge and leakage of cells out of the front due to random movement (cell diffusion). For illustration, consider the green dashed line indicating the density profile at a given earlier time. With only directed movement (along attractant gradient) and cell growth, the front peak would be higher at a later time, as illustrated by the brown dashed line. Instead, growth in the front bulge is compensated by back-diffusion of cells away from the front (green solid line indicating a later time). This compensation mechanism is further shown in the cylinder plots below, illustrating that the observed stable expansion dynamics (constant expansion speed, constant peak density) results from a dynamic balance between growth at the front and back-diffusion. The exponential density profile in the trailing region results from a combination of the steady outward movement of the front and a steady exponential growth of cells leaked out of the front¹⁴.

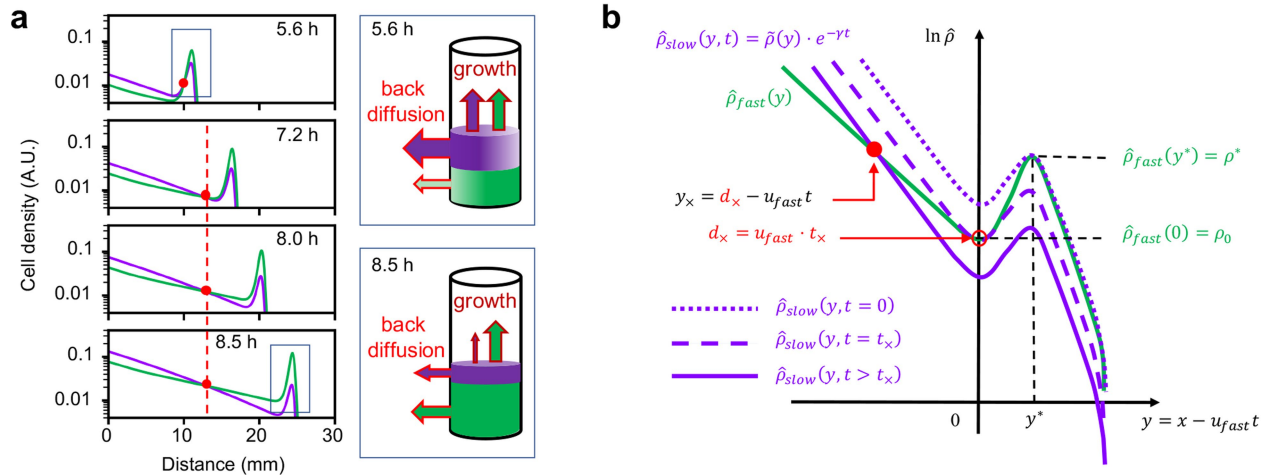


Extended Data Fig. 6 | Model of two-strain competitive expansion dynamics.

a, The GE model of bacterial range expansion¹⁴ (Extended Data Fig. 5) is extended to several strains of bacteria for which the densities are denoted by $\rho_i(x, t)$, with $i \in \{1, 2\}$ for two strains. The different bacterial strains are assumed to consume the same nutrient (n) and grow at the same rate $\lambda(n)$, in accordance with Monod's law. They also sense and consume the same signalling molecule (the attractant (a)) at the same rate $\mu(a)$. The random motion of each bacterial strain is described by a diffusion term, with effective diffusion coefficients D_i for strain i (equations (S5), (S6)). The spatial attractant profile $a(x, t)$ (resulting from bacterial consumption, equation (S8)) leads to directed motion which is modelled by a drift term in equations (S5) and (S6). The dependence of the drift velocities v_i on the attractant profile is given by equation (S7), which describes a range of proportional sensing $v \propto \partial_x a/a$ for $K_I < a < K_A$. The magnitude of the chemotactic response is parametrized by the chemotactic coefficient c_i for strain i . Finally, the dynamics of the nutrient and attractant are described by equations (S8) and (S9), with D_0 characterizing molecular diffusion.

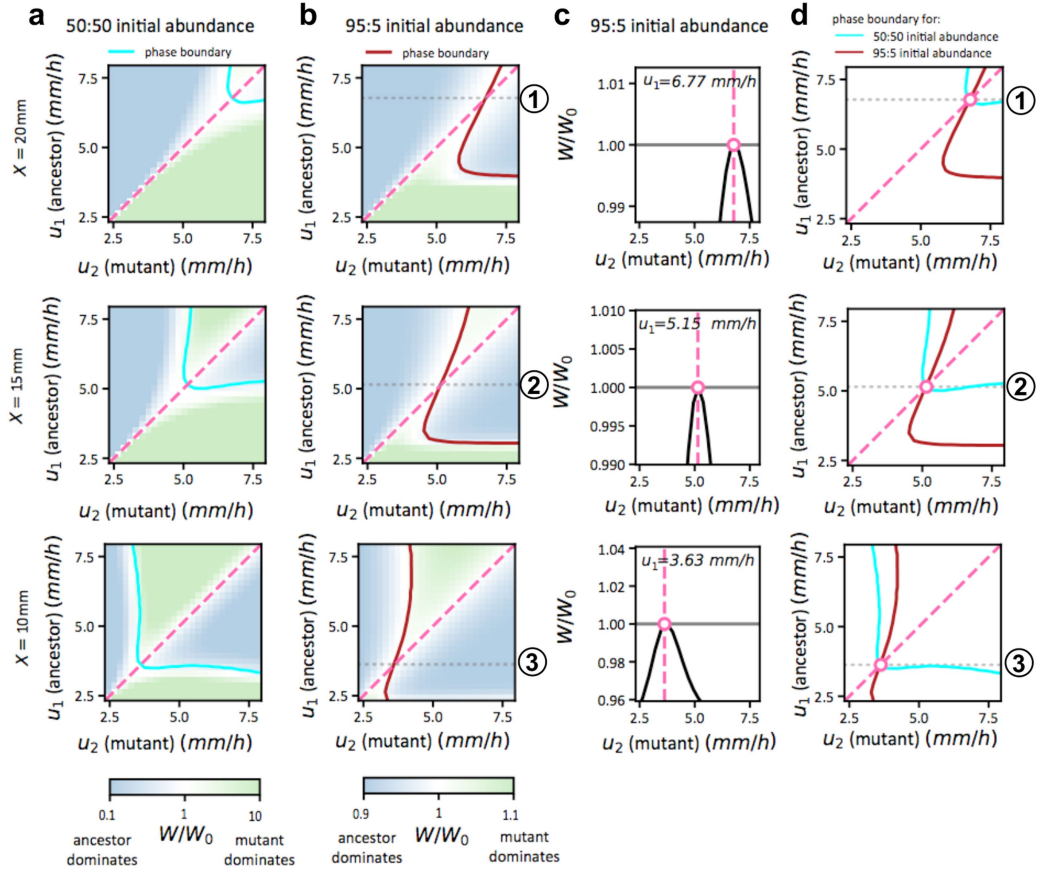
b, Outcomes of competitive expansion dynamics, showing the density profiles of two strains 24 h after inoculation with equal mixtures at the origin. Competition is run for one strain resembling the ancestor (black line, with expansion speed $u_{anc} = 6 \text{ mm h}^{-1}$) and the other strain resembling a mutant, with expansion speed u_{mut} increasing from left to right as indicated above the plots. Experimental data are from Extended Data Fig. 4l. **c**, Increase and abrupt freezing of the crossover distance over time as observed in simulations (line) and experiments (circles). The expansion speeds of the competing pair are u_{anc} versus $u_{mut} = 6 \text{ mm h}^{-1}$ versus 7 mm h^{-1} , with the latter mimicking the expansion speed of strain D (Extended Data Fig. 3b). Full temporal evolution of the density

profiles is shown in Supplementary Video 2. Experimental data are from Extended Data Fig. 4g. **d**, The chemotactic competition model in **a** is repeated to determine the stability distance d^* for various u and several smaller growth rate values λ . The results confirm that the linear relation in equation (2) between the stability distance and expansion speed shown in the main text holds for each growth rate simulated. **e**, The slope $u/d^*(u)$ in **d** is plotted against the growth rate λ . A linear dependence on λ is seen as predicted by Supplementary Analysis 2. **f**, Effect of differing growth rate (x axis) on the outcome of competition and crossover distance for two strains with different chemotactic coefficients (c_1 , c_2). The main result that the interior is dominated by the slower strain and the exterior by the faster strain holds. The crossover distance that separates the different regions of dominance varies smoothly with growth rate differences, but in opposite ways depending on whether the faster growing strain moves faster (red) or slower (green). The observed changes in crossover distances are minimal when growth-rate differences are minimal (for example, <5% as we observed experimentally for TB; Extended Data Fig. 1e) but become substantial when growth-rate differences become large (Supplementary Analysis 3). **g**, The stability distance d^* as in **d** but for an alternative model formulation of chemotactic movement⁵⁰ with different functions describing sensing and the directed movement along gradients. See Supplementary Analysis 4 for more details. Despite the model changes, a linear relation with growth-rate-dependent slopes was still observed. In **b–g**, the chemotactic coefficient c was varied to modify expansion speeds. The cellular diffusion coefficient D was varied accordingly, with $D = c/6.25$ remaining constant. Simulations with a fixed diffusion coefficient gave similar results. Experiments in **b**, **c** were repeated independently three times with similar results.



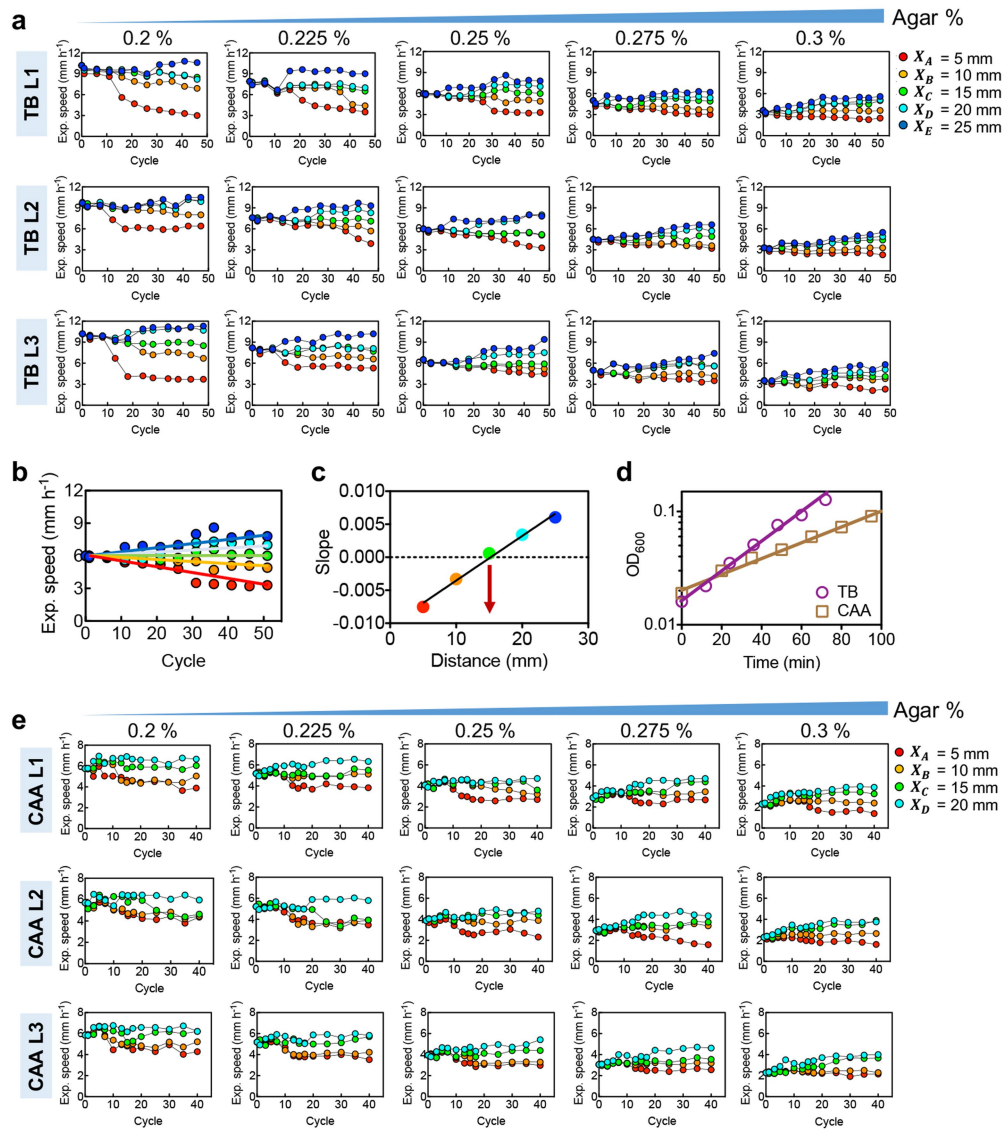
Extended Data Fig. 7 | Two-strain competition and crossover dynamics. The results in Fig. 2 show that competition between two strains with the same growth rate and different single-strain expansion speeds generically gives rise to a distinct spatial structure where the slower strain dominates inside and the faster strain dominates outside, separated by a crossover distance that depends on the two speeds. Here we explain how this feature arises from the underlying GE dynamics. **a**, Dynamics simulation using equations (S5)–(S9) for two strains differing only in their chemotactic coefficients ($c_{\text{slow}} = 652 \mu\text{m}^2 \text{s}^{-1}$ and $c_{\text{fast}} = 727 \mu\text{m}^2 \text{s}^{-1}$, with corresponding single-strain expansion speeds $u_{\text{slow}} = 4.95 \text{ mm h}^{-1}$ and $u_{\text{fast}} = 5.17 \text{ mm h}^{-1}$; other parameters, equal for both strains, are provided in Supplementary Table 3. The density profile of the strain with faster single-strain expansion speed is green, the other purple. Red circles indicate crossover distances, which freeze at a fixed position in the laboratory frame (dashed red line) after the troughs of the two density bulges cross. This feature (see also Extended Data Fig. 4g,i) allows us to define a simple crossover distance d_x that is time-invariant. To understand the crossover dynamics in **a**, we first provide a qualitative explanation. While the two strains would individually expand at different speeds, the competition dynamics involves a single co-migrating population. This is because the two strains chase after the same attractant profile, which can recede only at a single speed⁵¹. Because the front is moving faster than the speed at which the slower strain can stably propagate, the slower strain is gradually depleted from the front. As shown by the cylinders, at early times (before the front has reached the crossover distance and where the abundances of the two strains in the front bulge are comparable), the leakage flux of slower cells (purple) at the back exceeds that of the faster cells (green). As the two strains grow at the same rate behind the front, the slower strain dominates there. Because of its faster leakage, the slower strain is preferentially depleted from the front bulge. Subsequently, the leakage flux of the slower strain drops below that of the faster strain despite the faster leakage rate of the slower strain, owing to its reduced abundance. From there on, the back is dominated by the faster strain. At the crossover distance, the two leakage fluxes are the same. **b**, A simple analysis captures key features of the crossover dynamics and leads to a time-invariant crossover distance $d_x(u_{\text{fast}}, u_{\text{slow}})$ for two strains with single-strain expansion speeds u_{fast} and u_{slow} , and shows how this crossover distance leads to an evolutionarily stable selection distance d^* defined in the limit $u_{\text{slow}} \rightarrow u_{\text{fast}}$ (Figs. 3, 4). The two density profiles shown in **a**, $\rho_{\text{fast}}(x, t)$ and $\rho_{\text{slow}}(x, t)$ in the ‘laboratory frame’ coordinate x , are shown in **b** in the frame moving with speed u_{fast} . In this moving frame where the spatial coordinate is $y = x - u_{\text{fast}}t$, the density profile of the faster strain (green line) is approximately

stationary: $\rho_{\text{fast}}(x, t) = \hat{\rho}_{\text{fast}}(y)$ where $\hat{\rho}_{\text{fast}}(y)$ is the stationary solution of the single-strain dynamics in the reference frame moving at speed u_{fast} . Defining the position of the trough to be $y=0$ in the moving frame, $\hat{\rho}_{\text{fast}}(y)$ has a bulge for $y>0$ and a trailing exponential for $y<0$. In this frame, the density profile of the slower strain is generally not stationary, and is written as $\rho_{\text{slow}}(x, t) = \hat{\rho}_{\text{slow}}(y, t)$. As described in Supplementary Analysis 1, because the slower strain expands faster than its single-strain expansion speed its density bulge, while preserving the spatial structure, is depleted exponentially over time. This is expressed mathematically as $\hat{\rho}_{\text{slow}}(y, t) = \tilde{\rho}_{\text{slow}}(y)e^{-\gamma t}$ where $\tilde{\rho}_{\text{slow}}(y)$ also has a bulge for $y>0$ and a trailing exponential for $y<0$. Because the density bulges of the two strains are aligned spatially by the common attractant gradient, we denote the position of the bulge peak by y^* for both strains. However, in general neither the peak height nor the slope of the trailing exponentials would be the same between the two strains. In **b**, the density profile of the slower strain, $\hat{\rho}_{\text{slow}}(y)e^{-\gamma t}$, is illustrated at three times t . Dotted purple line indicates initial time ($t=0$) where the density bulge of the slower strain has the same peak value ρ^* as the faster strain (green line): $\tilde{\rho}_{\text{slow}}(y^*) = \rho^*$. Dashed purple line indicates density profile of the slower strain at time t_x , where the two density troughs cross: $\tilde{\rho}_{\text{slow}}(0)e^{-\gamma t_x} = \rho_0$. The location at which the troughs cross (open red circle) is the crossover distance $d_x = u_{\text{fast}}t_x$ (obtained from $y(t_x) = 0$). For $t>t_x$, the density bulge of the slower strain sinks steadily below that of the faster strain, and its density profile (solid purple line) crosses that of the faster strain (green line) at $y_x < 0$ behind the front and falls steadily backward in the moving frame (filled red circle). Supplementary Analysis 1 shows that for a steady sinking rate $\gamma \propto (u_{\text{fast}} - u_{\text{slow}})$, the corresponding crossover position in the laboratory frame remains at d_x for all $t > t_x$ (fixed position of red circle in **b**). This feature of the crossover dynamics can be understood intuitively: as chemotaxis occurs only within the bulge region, cells at the crossover point (and to its left) experience no net drift. As the two strains grow at the same rate, their densities at this position remain equal to each other for all $t > t_x$, implying that the density crossover is frozen in at d_x . The mathematical analysis provides an expression (Supplementary equation (E14); see Supplementary Information) of how the crossover distance $d_x = u_{\text{fast}}t_x$ depends on u_{fast} and u_{slow} and on static properties of the two density profiles, $\hat{\rho}_{\text{fast}}(y)$ and $\tilde{\rho}_{\text{slow}}(y)$. From this, we can obtain an expression for the stability distance, defined in the limit the two speeds approach each other: $d^*(u) = ut_x(u_{\text{slow}} \rightarrow u)$. Supplementary Analysis 2 shows that $t_x(u_{\text{slow}} \rightarrow u)$ approaches a finite limit that is proportional to $1/\lambda$, the doubling time. The proportionality of d^* and u is verified in Extended Data Fig. 6d, and the growth rate dependence of the slope in Extended Data Fig. 6e.



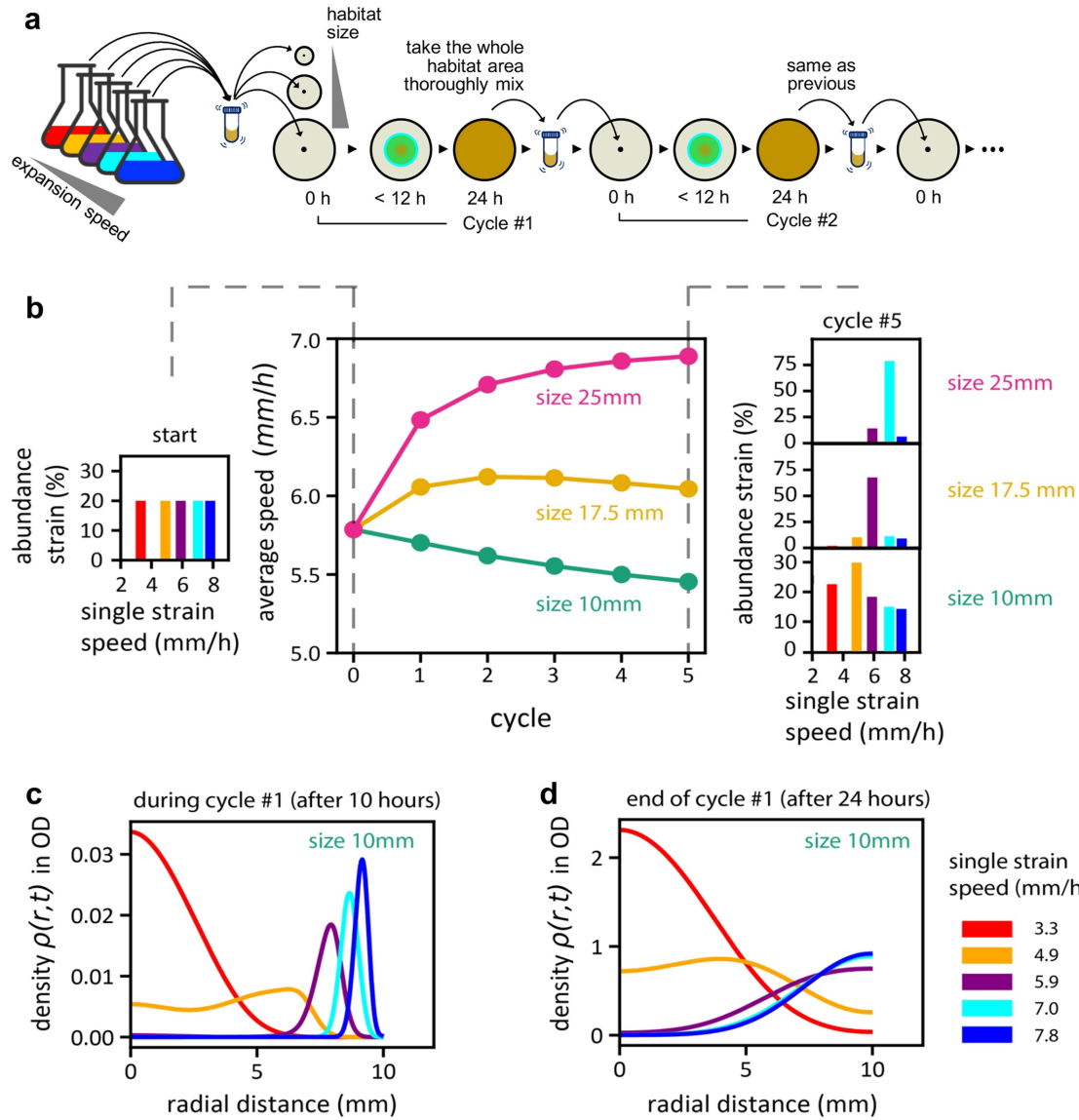
Extended Data Fig. 8 | Fitness landscapes for the competitive expansion process at different positions. The position-dependent fitness landscapes for the competition between two strains (strains 1 and 2, for example, an ancestor and a mutant), with expansion speeds u_1 and u_2 , respectively, computed according to the competitive expansion model in Extended Data Fig. 6a. The fitness of the mutant (strain 2) relative to the ancestor (strain 1) is defined as the ratio of its density with respect to the ancestor: $W(X) \equiv \lim_{t \rightarrow \infty} \left[\frac{\rho_2(X, t)}{\rho_1(X, t)} \right]$. In practice, we take $t = 24$ h, which is well after the dynamics have halted. We further normalize this fitness by the ratio of the initial inoculant, $W_0 \equiv \frac{\rho_2(X, 0)}{\rho_1(X, 0)}$, to obtain the ‘relative fitness gain’ W/W_0 . In **a**, **b**, green indicates a gain in the fitness of the mutant ($W/W_0 > 1$) and blue indicates a loss ($W/W_0 < 1$). Top, middle, and bottom rows refer to results at distances $X = 20, 15$ and 10 mm, respectively. **a**, Landscape W/W_0 for equal abundance (50:50) of the two strains at equal inoculation ($W_0 = 1$). Cyan and dashed represent $W/W_0 = 1$ and $u_1 = u_2$, respectively. For $W_0 = 1$, the cyan line is the crossover distance: $X = d_c(u, u')$ (corresponding to $W = 1$) (Fig. 4). The middle row ($X = 15$ mm) is an expanded view of Fig. 4c, except that the phase diagram in Fig. 4c shows only qualitative information on which strain dominates, whereas the fitness landscape here also provides quantitative information on the fitness gain. Top and bottom, corresponding fitness landscapes for $X = 20$ and 10 mm, respectively. **b**, Fitness landscape calculations for a mutant:ancestor inoculant ratio of 5:95 ($W_0 \equiv \rho_2(X, 0); \rho_1(X, 0) = 5:95$). The solid brown line still indicates the phase boundary $W/W_0 = 1$, but it is no longer directly related to the crossover distance

(see below). The topological features of the landscapes in **a** and **b** are similar but the details differ, reflecting the frequency-dependent nature of the competition process. **c**, Landscape profiles for three special ancestor expansion speeds, one at each position X indicated by dotted grey lines marked as ①, ② and ③ in **b**, each obtained as the intersection of the phase boundary and the diagonal (solid brown and dashed pink lines, respectively). For $X = 15$ mm (middle), ancestral speed $u_1^* = 5.15$ mm h⁻¹ marked by grey line ②, the fitness landscape has a maximum at $u_2 = u_1^*$ (white circle), indicating that ancestral expansion speed u_1^* is stable to invasion by mutants with smaller or larger expansion speeds. Similarly, for $X = 20$ mm (top), the ancestral speed $u_1^* = 6.77$ mm h⁻¹ (①) is the stable expansion speed, whereas for $X = 10$ mm (bottom), the ancestral speed $u_1^* = 3.63$ mm h⁻¹ (③) is the stable expansion speed. **d**, The phase boundaries shown as solid cyan and brown lines in **a**, **b** intersect the diagonal at the same speed u^* for each X . Thus, the stable expansion speeds can be obtained from the crossover distance (cyan line) based on 50:50 inoculant (equation (3)), despite the frequency dependence of the competition process. In other words, even though the overall fitness W depends on the initial frequency W_0 in a complex way, in the vicinity of the diagonal ($u_1 \approx u_2$), W is independent of W_0 so that the illustration in Fig. 3a of the stability of a strain of speed u at distance $d_c(u, u)$ is verified explicitly in **c**, **d**. To vary expansion speeds in each panel, the chemotactic coefficients c_i were varied within the range 180–1,500 $\mu\text{m}^2 \text{s}^{-1}$. Diffusion coefficients were changed accordingly ($D_i = c_i/6.25$; see Supplementary Model and Supplementary Table 3).



Extended Data Fig. 9 | Experimental evolution of expanding bacterial populations in various agar concentrations. **a**, Three lineages of evolution trajectories in semi-solid TB plates with different agar concentrations. **b**, **c**, Illustration of how stability distance is determined for each ancestral expansion speed based on the data shown in **a**. Using linear regression of the mean evolution trajectories obtained (here for 0.25% agar, ancestral ES = 6 mm h⁻¹, $n = 3$ biological replicates as in **a**), we obtained a slope for each series, A–E, as in **b**. Then we plotted the slope obtained in **b** against the

corresponding selection position to obtain another line (**c**). The stability distance is estimated as the x intercept of the line. **d**, Batch culture growth curve of ancestor cells grown in TB (purple) or CAA (brown) medium; the growth rate in CAA is almost 50% slower. Two replicates showed similar results. **e**, The evolution experiment in CAA medium for five agar densities (same as those shown for TB in **a**). Three independent lineages were run in parallel as for TB.



Extended Data Fig. 10 | Stable expansion speeds in spatial habitats of different sizes. a, Illustration of a recurring ecological scenario. A mixed community of species with different motility characteristics expands into an enclosed habitat of a certain size. All cells within the habitat have an equal chance of being passed on to occupy a new habitat of the same size. The competitive expansion model (Extended Data Fig. 6a) was extended to model the co-expansion of five distinct species, chemotaxing on the same attractant gradient (Supplementary Analysis 5). Each species has a distinct expansion speed when expanding alone (between 2 mm h^{-1} and 8 mm h^{-1}), due solely to different motility characteristics that are modelled by different chemotactic coefficients c_i in the range $\{34 \dots 274\} \mu\text{m}^2 \text{s}^{-1}$. A closed (zero-flux) boundary condition is applied so that when cells reach the edge of the habitat, they cannot propagate forward but they can propagate backward by diffusion (backward propagation via chemotaxis along reversed chemoattractant gradients is possible in principle but very limited, because the chemoattractant is mostly consumed when the population expands forward throughout the habitat). **b**, Average expansion speed of the population changes over different cycles of the simulated expansion–selection process. The average expansion speed increases as the selection proceeds for the

largest habitat size (25 mm, pink) and decreases for the smallest habitat size (10 mm, green). The relative abundance of species with different single-strain expansion speeds is changed after each round of selection. The result after the fifth round is shown on the right for the three habitat sizes shown. The faster species is enriched in the largest habitat (top) and the slower species is enriched in the smallest habitat (bottom). The slower species take more cycles to emerge as the winner since they occupy the habitat interior and receive lower weights owing to two-dimensional geometry. **c**, **d**, To understand why the faster species were selected against in the smaller habitat, we plotted the radial density function $\rho_i(r, t)$ for each species i at $t = 10 \text{ h}$ and 24 h after expansion during the first selection cycle (before any selection took place). The density profiles show that after the faster species reached the edge of the habitat, they moved backward towards the interior. However, the speed of this backward movement was limited compared to the outward movement. This is attributed to the fact that the chemoattractant behind the front is depleted, so the backward movement cannot rely on chemotaxis but is a result of cell diffusion (via Komolgorov–Fisher dynamics). In **d**, the cyan and blue lines (expansion speeds 7 and 7.8 mm h^{-1} , respectively) almost overlap and are hard to distinguish visually.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Imaging data was collected using the software NIS-Elements AR 4.50 (Nikon Ti-E Microsystems). The expansion speed data was collected using a Canon EOS 600D digital camera. The cell counting data was collected by software CytExpert2.1 (Beckman, Cyto-FLEX). The quantitative real-time RT-PCR data was collected by BIO-rad CFX and software Bio-rad CFX manager 3.1. Single cell tracking data was collected by Nikon Ti-E and MicroManager1.4. All the sequencing results were collected using a BGI Illumina system (HiSeq X10).

Data analysis

NIS-Elements AR 4.50 was used to analyze the fluorescence intensity profile data. Image J and MATLAB(2019a) were used to analyze the expansion speed data and motility data, CytExpert2.1 was used to analyze the cell counting data, Bio-rad CFX manager3.1 was used to analyze the quantitative real-time RT-PCR data. Sequencing results were analyzed with the breseq pipeline (breseq v0.31.1). GraphPad Prism6 was also used to analyze the growth rate and expansion speed. Custom-made codes used for performing simulations is available at https://github.com/jonascremer/chemotaxis_simulation.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All relevant data are included as source data and are available from the corresponding author on request. Sequencing data were deposited to the NCBI Sequence Read Archive(SRA) with the accession PRJNA559221. Data are available at <https://www.ncbi.nlm.nih.gov/sra/PRJNA559221>. Custom-made code used for performing simulations is available via GitHub at https://github.com/jonascremer/chemotaxis_simulation.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample size was determined according to the minimal number of independent biological replicates that significantly identified an effect.
Data exclusions	No data was excluded from the analysis.
Replication	Repeated measurements of the evolving quantities (expansion speeds, growth rates, relative fitness, profiles and mean run speed/ tumble bias) showed deviations of less than 10% confirming replication of the reported experiments.
Randomization	Not applicable.
Blinding	Data collection followed the same predetermined protocols throughout the whole study, so no blinding was performed. The custom made codes for the analysis of density profiles (to derive fitness) and the analysis of swimming behavior did not include any adjustable parameters and thus did not require blinding.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and

any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access and import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance

Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

State the source of each cell line used.

Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	<i>For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<i>Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."</i>
Recruitment	<i>Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.</i>
Ethics oversight	<i>Identify the organization(s) that approved the study protocol.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Samples were firstly fixed with pre-cooled cell counting buffer (0.9 % NaCl with 0.12 % formaldehyde). Then the fixed samples were diluted as necessary with staining buffer (cell counting buffer with 0.1 µg/ml DAPI) prior to the flow cytometer analysis. Details are provided in the Methods.

Instrument

Beckman, Cyto-FLEX

Software

CytExpert2.1

Cell population abundance

At least 50,000 cells per sample were analyzed.

Gating strategy

DAPI staining was applied to distinguish bacterial cells from other particles. DAPI-stained particles were deemed as the bacterial cells, and the purity was above 95%. Cells were separated into two groups GFP positive/GFP negative through the FITC channel. Obtained results agreed with fluorescence intensity measurements using microscopy and scanning the entire population.

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	<input type="checkbox"/> Used <input type="checkbox"/> Not used

Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	<input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both
Statistic type for inference (See Eklund et al. 2016)	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

Models & analysis

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>

NLRP3 inflammasome activation drives tau pathology

<https://doi.org/10.1038/s41586-019-1769-z>

Received: 14 March 2019

Accepted: 2 October 2019

Published online: 20 November 2019

Christina Ising^{1,2}, Carmen Venegas¹, Shuangshuang Zhang^{1,2}, Hannah Scheiblich^{1,2}, Susanne V. Schmidt³, Ana Vieira-Saecker^{1,2}, Stephanie Schwartz^{1,2}, Shadi Albasset^{1,2}, Róisín M. McManus^{1,2}, Dario Tejera², Angelika Griep², Francesco Santarelli², Frederic Brosseron², Sabine Opitz^{1,2}, James Stunden⁴, Maximilian Merten¹, Rakez Kaye⁵, Douglas T. Golenbock⁶, David Blum⁷, Eicke Latz^{2,3,6}, Luc Buée⁷ & Michael T. Heneka^{1,2,6*}

Alzheimer's disease is characterized by the accumulation of amyloid-beta in plaques, aggregation of hyperphosphorylated tau in neurofibrillary tangles and neuroinflammation, together resulting in neurodegeneration and cognitive decline¹. The NLRP3 inflammasome assembles inside of microglia on activation, leading to increased cleavage and activity of caspase-1 and downstream interleukin-1 β release². Although the NLRP3 inflammasome has been shown to be essential for the development and progression of amyloid-beta pathology in mice³, the precise effect on tau pathology remains unknown. Here we show that loss of NLRP3 inflammasome function reduced tau hyperphosphorylation and aggregation by regulating tau kinases and phosphatases. Tau activated the NLRP3 inflammasome and intracerebral injection of fibrillar amyloid-beta-containing brain homogenates induced tau pathology in an NLRP3-dependent manner. These data identify an important role of microglia and NLRP3 inflammasome activation in the pathogenesis of tauopathies and support the amyloid-cascade hypothesis in Alzheimer's disease, demonstrating that neurofibrillary tangles develop downstream of amyloid-beta-induced microglial activation.

Neuroinflammatory processes are critical in the development and progression of Alzheimer's disease¹. Amyloid-beta (A β) activates the NLRP3 inflammasome, which consists of NLRP3, ASC and caspase-1. Levels of active cleaved caspase-1 were elevated in amyloid-plaque-containing mice and patients with Alzheimer's disease^{4,5}, and knockout of ASC or NLRP3 ameliorated amyloid plaque pathology in transgenic doubly mutated APP/PS1 mice, a model for Alzheimer's disease³. As intracellular tau deposits are not the predominant pathological signature, Alzheimer's disease is considered to be a secondary tauopathy⁶. Primary tauopathies, such as frontotemporal dementia (FTD), also present with neuroinflammation and cognitive deficits, which can be modelled in mice by overexpressing mutated human tau.

NLRP3 inflammasome activity in FTD

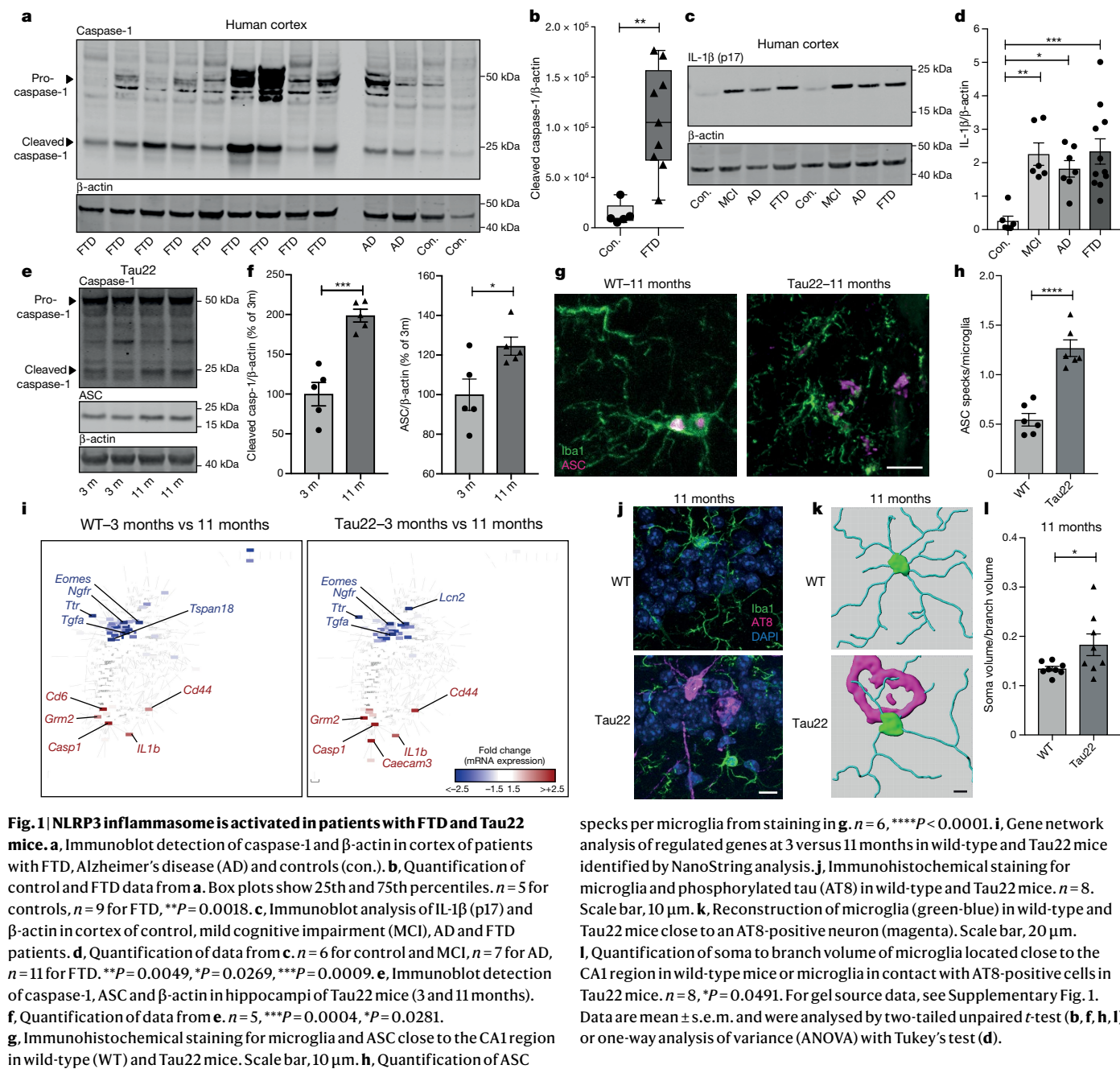
To identify a potential role of the NLRP3 inflammasome in patients with tauopathy, we analysed cortex samples from patients with FTD. We found elevated cleavage of caspase-1 and increased ASC levels and mature IL-1 β (p17) (Fig. 1a–d, Extended Data Fig. 1a–c), indicative of NLRP3 inflammasome activation. As a model, we used Tau22 mice, which express human tau FTD mutations and develop tau pathology over time⁷. We confirmed NLRP3 inflammasome activation by detection of increased levels of cleaved caspase-1 as well as ASC in cerebral

samples from 11-month-old compared with 3-month-old Tau22 mice (Fig. 1e, f). In addition, we detected significantly more extracellular ASC specks, which are released upon NLRP3 inflammasome activation, in Tau22 mice by immunohistochemistry (Fig. 1g, h, Extended Data Fig. 1d) and increased levels of cleaved caspase-1 and IL-1 β (p17) when compared with wild-type mice (Extended Data Fig. 1e–h).

Gene expression analysis of neuroinflammation-associated genes of cerebral samples (Extended Data Fig. 2a) revealed age- but not genotype-related differences between wild-type and Tau22 mice (Extended Data Fig. 2b). We visualized the dynamics of gene expression patterns in a co-regulation network (Fig. 1i, Extended Data Fig. 2c). *Casp1* and *Il1b* were amongst the top upregulated genes, indicating that the NLRP3 inflammasome pathway is induced and possibly has a central role in normal ageing, as previously reported⁸, as well as neuroinflammatory and degenerative processes. Self-organizing map (SOM) clustering identified genes participating in tau pathogenicity. Genes were grouped according to similar expression levels at indicated time points in wild-type or Tau22 mice (Extended Data Fig. 2d) and defined as signature genes for each condition on the basis of elevated expression levels (Supplementary Table 1). Gene set enrichment analysis (GSEA) of highly expressed genes in 3-month-old wild-type animals were enriched in biological processes such as 'translation' or 'synthetic processes' (gene signature (i), Extended Data Fig. 2d). Microglia programming changed

¹Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital of Bonn, Bonn, Germany. ²German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany.

³Institute of Innate Immunity, University Hospital Bonn, Bonn, Germany. ⁴IFM Therapeutics GmbH, Bonn, Germany. ⁵Mitchell Center for Neurodegenerative Diseases and Departments of Neurology, Neuroscience and Cell Biology, University of Texas Medical Branch, Galveston, TX, USA. ⁶Division of Infectious Diseases and Immunology, University of Massachusetts Medical School, Worcester, MA, USA. ⁷University of Lille, Inserm, CHU-Lille, UMR-S 1172, "Alzheimer & Tauopathies", Labex DISTALZ, Lille, France. *e-mail: michael.heneka@ukbonn.de



during ageing (8 months) indicated by highly expressed genes associated with 'trans-synaptic signalling' and 'regulation of communication' (data not shown). However, regulation of these gene clusters was absent in Tau22 mice. The reduced number of genes in gene signature (iii) prohibited a GSEA. Genes in gene signature (iv) in 3-month-old Tau22 mice revealed an association with immune responses, indicating an involvement of immunological processes at disease onset. At later stages of disease development, genes upregulated in Tau22 mice participated in functions such as 'response to stress' (gene signature (vi), Extended Data Fig. 2e). GSEA of genes that are highly expressed in 8-month-old Tau22 mice was not feasible owing to low numbers of genes (Supplementary Table 1). Previous studies have described an active type I interferon (IFN) signature in Alzheimer's disease⁹. Mice lacking IFNAR1 were resistant to A β ₁₋₄₂-induced neurotoxicity¹⁰. Up to 73% of genes from gene signatures (iv) and (vi) are associated with interferon (Extended Data Fig. 2f). We visualized central highly interacting genes in a network, which we hypothesized might regulate or

collaborate with other genes during disease development. The network for genes upregulated in Tau22 mice at three months showed two clusters of highly interactive genes (Extended Data Fig. 3a): one cluster associated with pro-inflammatory signalling cascades such as *Jun*, *Fas* and Toll-like receptors; and a second cluster consisting of chromatin remodellers, such as *Hdac2*, indicating that the function of microglia at early stages of tau pathology involves remodelling of the epigenetic landscape. The network for genes with higher expression in Tau22 mice at 11 months was less dense (Extended Data Fig. 3b). Central genes were *Irf1*, *MyD88* and *Il1rap*, which indicate the inflammatory phenotype of tau pathology. We next studied microglia and astrocyte morphology close to hyperphosphorylated tau-positive neurons (indicated by AT8 staining) in the CA1 region at different stages of tau pathology (3, 8 and 11 months of age). This analysis revealed microglial changes at 8 and 11 months when substantial tau pathology is present in these mice (Fig. 1j–l, Extended Data Fig. 4a), whereas astrocyte morphology did not change (Extended Data Fig. 4b).

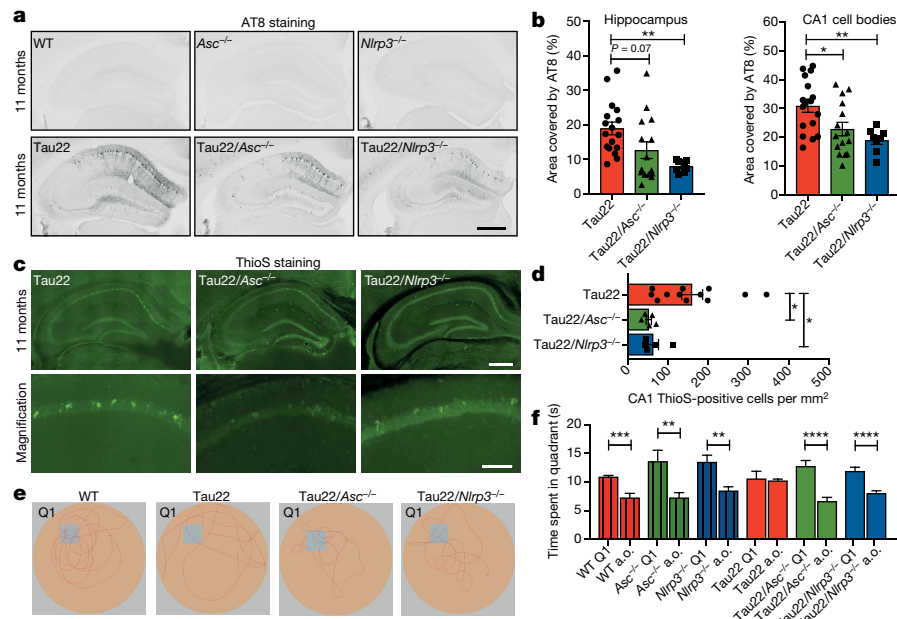


Fig. 2 | Loss of NLRP3 inflammasome function decreases tau pathology and prevents cognitive decline. **a**, Immunohistochemical staining for phosphorylated tau (AT8) in mouse hippocampi. Scale bar, 500 μ m. **b**, Quantification of AT8 in hippocampus and CA1 region shown in **a**. $n = 17$ for Tau22, $n = 15$ for Tau22/*Asc*^{-/-}, $n = 8$ for Tau22/*Nlrp3*^{-/-}. Hippocampus: $**P = 0.0055$, CA1: $*P = 0.0278$, $**P = 0.0059$. **c**, Staining with thioflavine S (thioS) (aggregated tau) of mouse hippocampi. Scale bars, 500 μ m (top) and 100 μ m (bottom). **d**, Quantification of thioflavine-S-positive cells in CA1 region

shown in **c**. $n = 12$ for Tau22, $n = 5$ for Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-}. Tau22 versus Tau22/*Asc*^{-/-}: $*P = 0.0240$, Tau22 versus Tau22/*Nlrp3*^{-/-}: $*P = 0.0444$. **e**, Example of movement of mice at probe trial day 9 of the Morris water maze test. **f**, Quantification of time spent in quadrant 1 (Q1) versus all other quadrants (a.o.) at probe trial day 9. $n = 12$ for wild-type, *Nlrp3*^{-/-}, Tau22/*Asc*^{-/-}, Tau22/*Nlrp3*^{-/-}, $n = 14$ for *Asc*^{-/-}, $n = 16$ for Tau22. $***P = 0.0001$, $****P < 0.0001$, *Asc*^{-/-}: $**P = 0.0065$, *Nlrp3*^{-/-}: $**P = 0.0012$. Data are mean \pm s.e.m. and were analysed by one-way ANOVA with Tukey's test (**b**, **d**) or two-tailed unpaired *t*-test (**f**).

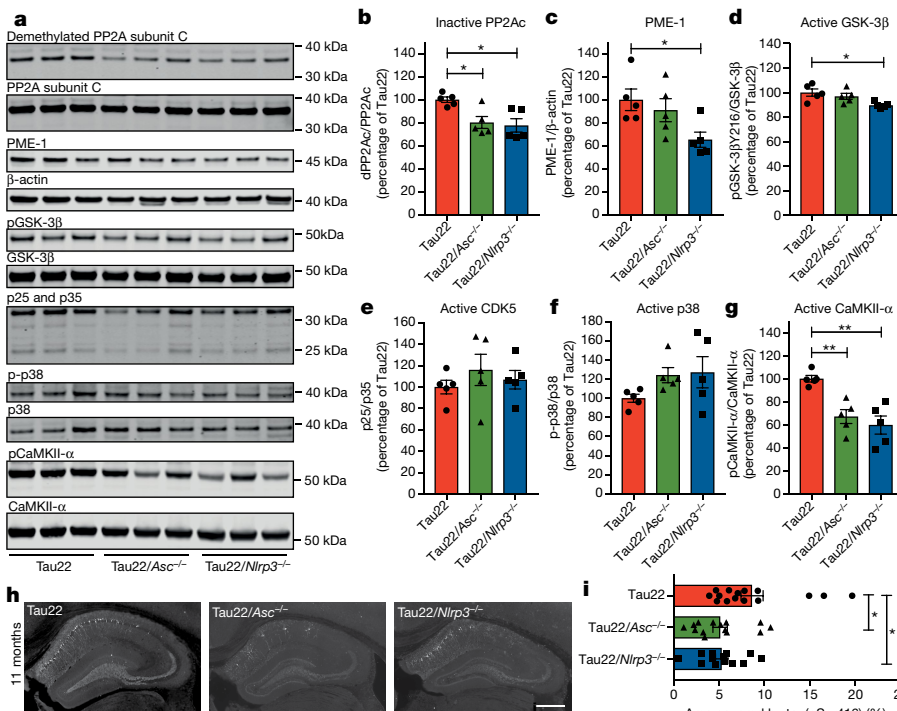


Fig. 3 | *Nlrp3*- and *Asc*-knockout inhibits CaMKII- α and promotes phosphatase activity. **a**, Immunoblot analysis of mouse hippocampi (11 months) stained for demethylated PP2A subunit C, total PP2A subunit C, PP2A methylesterase (PME-1), β -actin, GSK-3 β phosphorylated at Tyr216 (pGSK-3 β), total GSK-3 β , p25/p35, phosphorylated p38 (p-p38), total p38, calmodulin-dependent protein kinase II α phosphorylated at Thr286 (pCaMKII- α) and total CaMKII- α . **b–g**, Quantification of the enzyme activities or abundance shown in **a**. $n = 5$. PP2Ac: Tau22 versus Tau22/*Asc*^{-/-}: $*P = 0.0286$, Tau22 versus

Tau22/*Nlrp3*^{-/-}: $*P = 0.0144$, PME1: $*P = 0.0398$, GSK-3 β : $*P = 0.0205$, CaMKII- α : Tau22 versus Tau22/*Asc*^{-/-}: $**P = 0.0055$, CaMKII- α Tau22 versus Tau22/*Nlrp3*^{-/-}: $**P = 0.0012$. **h**, Immunohistochemical staining for tau phosphorylated at Ser416 (Tau-pSer416) in mouse hippocampi. Scale bar, 500 μ m. **i**, Quantification of Tau-pSer416 in CA1 region shown in **h**. $n = 15$ for Tau22, $n = 14$ for Tau22/*Asc*^{-/-}, $n = 13$ for Tau22/*Nlrp3*^{-/-}. Tau22 versus Tau22/*Asc*^{-/-}: $*P = 0.0314$, Tau22 versus Tau22/*Nlrp3*^{-/-}: $*P = 0.0476$. For gel source data, see Supplementary Fig. 1. Data are mean \pm s.e.m. and were analysed by one-way ANOVA with Tukey's test.

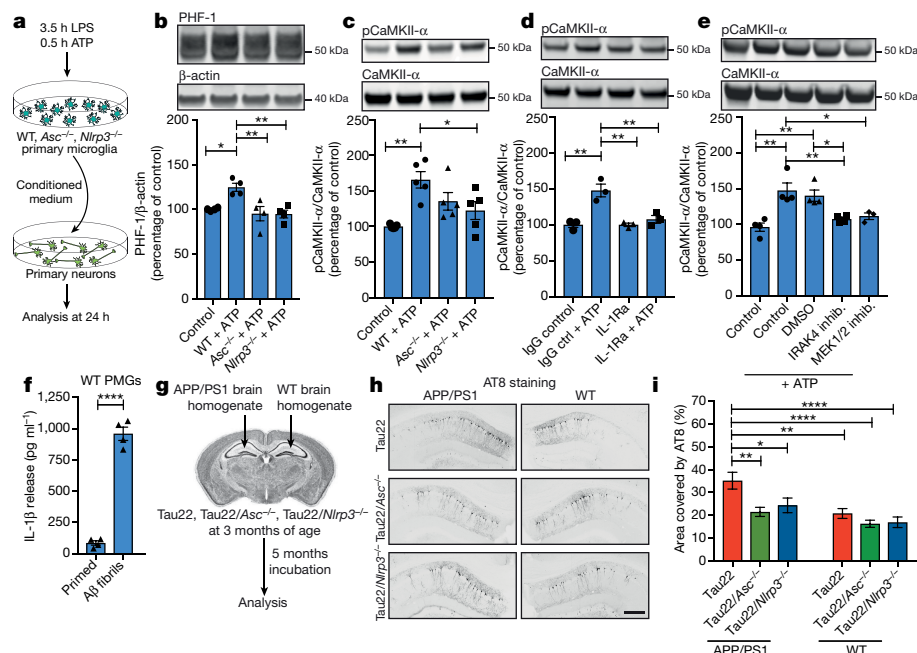


Fig. 4 | Inflammasome function is necessary for Aβ-induced tau pathology.

a, Schematic illustrating the experimental set-up used for experiments shown in **b–e**. **b**, **c**, Immunoblot analysis and quantification of tau phosphorylated at Ser396 and Ser404 (PHF-1) and pCaMKII-α in treated neurons (primary wild-type microglia (control), LPS- and ATP-activated wild-type microglia (wild type + ATP), LPS- and ATP-activated *Asc*^{-/-} or *Nlrp3*^{-/-} knockout microglia (*Asc*^{-/-} + ATP or *Nlrp3*^{-/-} + ATP)). *n* = 4 in **b**, with control versus wild type + ATP: *P* = 0.0235, wild type + ATP versus *Asc*^{-/-} + ATP: *P* = 0.0072, wild type + ATP versus *Nlrp3*^{-/-} + ATP: *P* = 0.0064. *n* = 5 in **c** with *P* = 0.0022, *P* = 0.0454. **d**, Immunoblot analysis and quantification of pCaMKII-α in neurons treated with an IL-1 receptor antagonist (IL-1Ra) or the corresponding isotype control in addition to conditioned medium from wild-type microglia. *n* = 3, IgG control + ATP versus IgG control and versus IL-1Ra: *P* = 0.0015, IgG control + ATP versus IL-1Ra + ATP: *P* = 0.0044. **e**, Immunoblot analysis and quantification of pCaMKII-α in neurons after treatment with DMSO, IRAK4 inhibitor (inhib.) PF06650833 or MEK1/2 inhibitor UO126 in addition to conditioned medium

from wild-type microglia. *n* = 3 for MEK1/2 inhibitor, *n* = 4 for all other groups, control versus control + ATP: *P* = 0.0011, control versus DMSO + ATP: *P* = 0.0040, control + ATP versus IRAK4 inhibitor + ATP: *P* = 0.0087, control + ATP versus MEK1/2 inhibitor + ATP: *P* = 0.0309, DMSO + ATP versus IRAK4 inhibitor: *P* = 0.0338. **f**, IL-1β levels in conditioned medium of primary wild-type microglia treated with LPS and 10 μM Aβ fibrils. *n* = 4, *P* < 0.0001. **g**, Schematic for injection model. **h**, Immunohistochemical staining for phosphorylated tau (AT8) of CA1 region of mice injected with either APP/PS1 or wild-type brain homogenates. Scale bar, 250 μm. **i**, Quantification of AT8 in CA1 region of injected mice shown in **h**. *n* = 18 sections of *n* = 6 mice for Tau22, *n* = 30 sections of *n* = 6 mice for Tau22/*Asc*^{-/-}, *n* = 25 sections of *n* = 6 mice for Tau22/*Nlrp3*^{-/-}. *P* = 0.0490, *P* < 0.0001, Tau22 + APP/PS1 versus Tau22/*Asc*^{-/-} + APP/PS1: *P* = 0.0034, Tau22 + APP/PS1 versus Tau22 + wild type: *P* = 0.0071. For gel source data, see Supplementary Fig. 1. Data are mean ± s.e.m. and were analysed by one-way ANOVA with Tukey's test (**b–e**, **i**) or two-tailed unpaired *t*-test (**f**).

NLRP3 loss protects from tau pathology

To assess whether the NLRP3 inflammasome is involved in the pathogenesis of tauopathies, we crossed Tau22 mice with mice that are deficient in *Pycard* (Tau22/*Asc*^{-/-}) or *Cias1* (Tau22/*Nlrp3*^{-/-}). Notably, we detected lower levels of cleaved caspase-1 and IL-1β as well as reduced ASC speck formation and release in NLRP3-inflammasome-deficient mice (Extended Data Fig. 5a–e), thereby confirming that ASC and NLRP3 mediate activation of the inflammasome in Tau22 mice. Analysis of tau (AT8) revealed lower levels of tau hyperphosphorylation in the hippocampus, CA1 cell body region and granular cell layer of the dentate gyrus in aged Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} mice (Fig. 2a, b, Extended Data Fig. 6a–d). Additionally, aggregated tau levels, as assessed by thioflavin staining, were reduced at 11 months (Fig. 2c, d) and less misfolded tau—as indicated by MC1 analysis—was present in the soluble fractions of hippocampus of aged Tau22 mice deficient for NLRP3. Human tau levels were decreased in Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} mice at 8 months, but not 11 months of age (Extended Data Fig. 6e–h), probably indicating a plateau in tau expression at this age⁷. Furthermore, the loss of NLRP3 inflammasome components rescued the spatial memory deficits present in Tau22 mice⁷ (Fig. 2e, f). Our analysis of microglia numbers and area covered by astrocytes at 8 and 11 months of age showed no significant differences between genotypes (Extended Data Fig. 7a–d).

NLRP3 regulates kinases and phosphatases

Several kinases and one major phosphatase (PP2A) regulate tau phosphorylation¹¹. We found decreased levels of the inactive phosphatase PP2A accompanied by lower levels of its negative regulator PME-1¹² in hippocampus samples of Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} mice (Fig. 3a–c). Furthermore, GSK-3β kinase activity was reduced in Tau22/*Nlrp3*^{-/-} mice, whereas the kinase activities of CDK5 (measured by the ratio of its regulators p25 and p35) and p38 remained unchanged (Fig. 3d–f). By contrast, active CaMKII-α was diminished in both Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} mice (Fig. 3g), probably explaining the reduced levels of tau phosphorylation at serine 416 (Ser416; a CaMKII-α target site) in the CA1 region of Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} (Fig. 3h, i). To identify inflammatory genes that could be involved in the reduction of tau pathology, we compared the gene expression of cerebral samples from Tau22 and Tau22/*Nlrp3*^{-/-} mice (Extended Data Fig. 8a). We identified 18 genes that were either significantly induced or suppressed in Tau22/*Nlrp3*^{-/-} mice (Extended Data Fig. 8b–d). Among these, *Ccl3* was reduced in 11-month-old Tau22/*Nlrp3*^{-/-} mice, which represents a microglia-produced pro-inflammatory component, and its reduction improves memory in Tau22 mice¹³. The immunoreceptor CD300lf, which is expressed in astrocytes, oligodendrocytes and microglia, was downregulated in Tau22 and upregulated in Tau22/*Nlrp3*^{-/-} mice at all ages. Overexpression of CD300lf has previously been shown to be neuroprotective in a model of acute brain injury¹⁴. In addition, ARC,

a protein implicated in synaptic plasticity and memory formation¹⁵, was upregulated in aged Tau22/*Nlrp3*^{-/-} mice, suggesting a close interaction between microglia and neurons. We found no evidence of gut inflammation in Tau22 and Tau22/*Nlrp3*^{-/-} mice (Extended Data Fig. 9a–o).

To study the regulation of tau kinases and phosphatases in more detail, we validated key *in vivo* findings by performing additional *in vitro* analysis. Conditioned medium was collected from primary microglia generated from wild-type, *Asc*^{-/-} and *Nlrp3*^{-/-} mice after LPS priming and ATP activation to induce the NLRP3 inflammasome and subsequently added to primary mouse hippocampal neuron cultures (Fig. 4a). Treatment with conditioned medium from wild-type, but not *Asc*^{-/-} or *Nlrp3*^{-/-}, microglia resulted in increased levels of mouse tau phosphorylated at Ser396/Ser404 (PHF-1) as well as augmented total mouse tau levels and higher levels of active CaMKII- α (Fig. 4b, c, Extended Data Fig. 10a). Inhibition of the neuronal IL-1 receptor or their downstream effectors IRAK4 and MEK1/2 in the IL-1 β signalling pathway antagonized the effects on CaMKII- α (Fig. 4d, e), indicating that microglia-derived IL-1 β worsens tau pathology, as observed with other tau kinases¹⁶.

Other amyloidogenic proteins such as A β can activate the NLRP3 inflammasome⁴. Similarly, tau can also activate the NLRP3 inflammasome resulting in IL-1 β release from microglia, as evidenced by treatments with brain homogenates from Tau22—but not wild-type—mice (Extended Data Fig. 10b). To determine the tau species involved, we used human wild-type tau and P301S-mutated tau in either their monomeric, oligomeric or fibrillar forms. Tau monomers and oligomers significantly increased IL-1 β secretion in an ASC- and NLRP3-dependent manner. However, in contrast to a previous report¹⁷, tau fibrils showed only small, non-significant effects (Extended Data Fig. 10c, d). The effect mediated by tau monomers could also be blocked by using a pharmacological specific NLRP3 inhibitor, CRID3 (Extended Data Fig. 10e, f). Treatment with tau monomers and oligomers also resulted in the release of cleaved caspase-1 into the cell supernatants (Extended Data Fig. 10g–j).

NLRP3 mediates A β -induced tau pathology

A role for microglia in tau seeding and spreading has previously been suggested^{17,18}. In Alzheimer's disease pathology, amyloid plaque formation precedes tau pathology. Of note, tau pathology can also be induced by intracerebral injection of A β fibrils or A β -containing brain homogenates into tau-transgenic mice by as yet unknown mechanisms^{19,20}. To investigate the effect of NLRP3 inflammasome activation on A β -induced tau seeding, we first demonstrated that A β fibrils activate the NLRP3 inflammasome as indicated by IL-1 β release in culture (Fig. 4f). We next injected brain homogenates from APP/PS1 or wild-type mice into the hippocampus of inflammasome-knockout mice (Fig. 4g). Injection of APP/PS1 brain homogenate efficiently induced tau hyperphosphorylation in the CA1 region in Tau22 but not Tau22/*Asc*^{-/-} or Tau22/*Nlrp3*^{-/-} mice (Fig. 4h, i), suggesting that NLRP3 activity represents an essential component in the A β -tau cascade.

In summary, this work places NLRP3 activation upstream of tau pathology in Tau22 mice. NLRP3 activation induces tau hyperphosphorylation and aggregation, at least partially through tau kinases in an IL-1 β -dependent manner. Tau oligomers and fibrils, but not monomers, have previously been shown to promote tau pathology and neuronal degeneration in assays with direct application of tau species on neuronal cultures^{21,22}. Here we provide evidence that tau oligomers and monomers have direct effects on microglia by activating NLRP3.

As non-fibrillar tau can actively be released by neurons^{23,24}, it could thereby contribute to chronic microglial activation in tauopathies. This pathway could potentially be blocked by NLRP3 inhibitors or by inhibition of tau binding to microglia. Furthermore, NLRP3 activation mediates A β -induced tau pathology in Tau22 mice, suggesting that patients with Alzheimer's disease may potentially benefit from NLRP3-directed treatment strategies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1769-z>.

1. Ising, C. & Heneka, M. T. Functional and structural damage of neurons by innate immune mechanisms during neurodegeneration. *Cell Death Dis.* **9**, 120 (2018).
2. Heneka, M. T., McManus, R. M. & Latz, E. Inflammasome signalling in brain function and neurodegenerative disease. *Nat. Rev. Neurosci.* **19**, 610–621 (2018).
3. Venegas, C. et al. Microglia-derived ASC specks cross-seed amyloid- β in Alzheimer's disease. *Nature* **552**, 355–361 (2017).
4. Halle, A. et al. The NALP3 inflammasome is involved in the innate immune response to amyloid-beta. *Nat. Immunol.* **9**, 857–865 (2008).
5. Heneka, M. T. et al. NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice. *Nature* **493**, 674–678 (2013).
6. Lewis, J. & Dickson, D. W. Propagation of tau pathology: hypotheses, discoveries, and yet unresolved questions from experimental and human brain studies. *Acta Neuropathol.* **131**, 27–48 (2016).
7. Schindowski, K. et al. Alzheimer's disease-like tau neuropathology leads to memory deficits and loss of functional synapses in a novel mutated tau transgenic mouse without any motor deficits. *Am. J. Pathol.* **169**, 599–616 (2006).
8. Youm, Y.-H. et al. Canonical Nlrp3 inflammasome links systemic low-grade inflammation to functional decline in aging. *Cell Metab.* **18**, 519–532 (2013).
9. Taylor, J. M. et al. Type-1 interferon signaling mediates neuro-inflammatory events in models of Alzheimer's disease. *Neurobiol. Aging* **35**, 1012–1023 (2014).
10. Minter, M. R. et al. Soluble amyloid triggers a myeloid differentiation factor 88 and interferon regulatory factor 7 dependent neuronal type-1 interferon response *in vitro*. *J. Neuroinflammation* **12**, 71 (2015).
11. Iqbal, K. et al. Tau pathology in Alzheimer disease and other tauopathies. *Biochim. Biophys. Acta* **1739**, 198–210 (2005).
12. Ortega-Gutiérrez, S., Leung, D., Ficarro, S., Peters, E. C. & Cravatt, B. F. Targeted disruption of the PME-1 gene causes loss of demethylated PP2A and perinatal lethality in mice. *PLoS ONE* **3**, e2486 (2008).
13. Laurent, C. et al. Hippocampal T cell infiltration promotes neuroinflammation and cognitive decline in a mouse model of tauopathy. *Brain J. Neurol.* **140**, 184–200 (2017).
14. Peluffo, H. et al. Overexpression of the immunoreceptor CD300f has a neuroprotective role in a model of acute brain injury. *Brain Pathol.* **22**, 318–328 (2012).
15. Epstein, I. & Finkbeiner, S. The Arc of cognition: Signaling cascades regulating Arc and implications for cognitive function and disease. *Semin. Cell Dev. Biol.* **77**, 63–72 (2018).
16. Bhaskar, K. et al. Regulation of tau pathology by the microglial fractalkine receptor. *Neuron* **68**, 19–31 (2010).
17. Stancu, I.-C. et al. Aggregated Tau activates NLRP3-ASC inflammasome exacerbating exogenously seeded and non-exogenously seeded Tau pathology *in vivo*. *Acta Neuropathol.* **137**, 599–617 (2019).
18. Asai, H. et al. Depletion of microglia and inhibition of exosome synthesis halt tau propagation. *Nat. Neurosci.* **18**, 1584–1593 (2015).
19. Götz, J., Chen, F., van Dorpe, J. & Nitsch, R. M. Formation of neurofibrillary tangles in P301L tau transgenic mice induced by A β 42 fibrils. *Science* **293**, 1491–1495 (2001).
20. Bolmont, T. et al. Induction of tau pathology by intracerebral infusion of amyloid-beta-containing brain extract and by amyloid-beta deposition in APP \times Tau transgenic mice. *Am. J. Pathol.* **171**, 2012–2020 (2007).
21. Shafiei, S. S., Guerrero-Muñoz, M. J. & Castillo-Carranza, D. L. Tau oligomers: cytotoxicity, propagation, and mitochondrial damage. *Front. Aging Neurosci.* **9**, 83 (2017).
22. Usenovic, M. et al. Internalized tau oligomers cause neurodegeneration by inducing accumulation of pathogenic tau in human neurons derived from induced pluripotent stem cells. *J. Neurosci. Off. J. Soc. Neurosci.* **35**, 14234–14250 (2015).
23. Karch, C. M., Jeng, A. T. & Goate, A. M. Extracellular Tau levels are influenced by variability in Tau that is associated with tauopathies. *J. Biol. Chem.* **287**, 42751–42762 (2012).
24. Yamada, K. et al. Neuronal activity regulates extracellular tau *in vivo*. *J. Exp. Med.* **211**, 387–393 (2014).

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Animal procedures and sample collection

All mice were on a C57BL/6N genetic background. THY-Tau22 transgenic mice (named Tau22) express a four-repeat isoform of human tau (1N4R) with a G272V and P301S mutation under control of the Thy1.2 promoter. These mice were previously generated and characterized⁷ and were used in a heterozygous state in this study. Tau22 mice were crossed with either *Pycard*-knockout mice (named *Asc*^{-/-}; Millennium Pharmaceuticals) or *Cias1*-knockout mice (named *Nlrp3*^{-/-}; Millennium Pharmaceuticals)²⁵. Non-tau-transgenic littermates were used as controls throughout the study. Mice were housed according to the standardized conditions in the University Hospital of Bonn animal facility. All mouse studies complied with relevant ethical regulations and were performed as approved by the local ethical committee (LANUV NRW 84-02.04.2017.A226). As no overt difference in tau pathology between aged male and female Tau22 mice is reported⁷, mixed genders were used throughout the study. Power analysis was used to predetermine the sample size. Mice were grouped according to genotype before random assignment to the experimental conduct. Researchers performing animal experiments were blinded to the genotype.

At 3, 8 and 11 months of age or 5 months after stereotaxic surgery, samples were collected. Blood was taken for serum analysis and the mice were transcardially perfused with 1× PBS. Brains were taken out and the hemispheres either separated or kept in one piece. Brain samples assigned for immunohistochemical staining were immersion fixed in 4% PFA/1× PBS for 24 h at 4 °C before sectioning on a Leica Vibratome (40 µm thickness). Sections were kept in 1× PBS supplemented with sodium azide at 4 °C. The other brain samples were dissected into cortex and hippocampus before freezing on dry ice and storage at -80 °C. The caecum and the colon were dissected free, the colon length was assessed and the caecum was weighed. The spleen, part of the medial intestine and the colon were snap-frozen in liquid nitrogen for further analysis. Faecal samples were taken from the distal end of the colon and snap-frozen in liquid nitrogen.

Antibodies and reagents

For tissue stainings, anti-biotinylated AT8 (Thermo Fisher Scientific, MN1020B, lot SB2334646, 1:500), anti-IBA1 (Wako, 019-19741, lot PTE055, 1:400 and Abcam, ab5076, lot GR3245261-1, 1:1500), anti-GFAP (Invitrogen, 13-0300, lot SA247423, 1:100) and anti-pTau-S416 (Abcam, ab119391, lot GR93695-21, 1:1,000) were used. For immunoblot analysis, the following antibodies were used: anti-Tau5 (Thermo Fisher Scientific, MA5-12808, lot sc2348223, 1:500), anti-MC1 (gift from P. Davies, 1:1,000), anti-PHF-1 (gift from P. Davies, 1:1,000), anti-caspase-1 (Genentech, clone 4B4.2.1, gift from Genentech, 1:1,000; Adipogen, clone Bally-1, AG-20B-0048, lot 26101409, 1:1,000; Adipogen, clone Casper-1, AG-20B-0042, lot A28881708, 1:50), anti-IL-1β (Gene Tex, GTX74034, lot 42900, 1:1,000), anti-ASC (Adipogen, clone Alz177, AG-25B-0006, lot A40221902, 1:1,000), anti-β-actin (Cell Signaling, 4967, lot 11, 1:2,000), anti-pCaMKIIα (Cell Signaling, clone D21E4, 12716T, 1:1,000), anti-CaMKIIα (Cell Signaling, clone 6G9, 50049S, lot 1, 1:1,000), anti-pGSK-3β (BD, clone 13A, 612313, lot 3768, 1:1,000), anti-GSK-3β (Cell Signaling, clone 27C10, 9315S, lot 14, 1:1,000), anti-p25/p35 (Cell Signaling, clone C64B10, 2680S, lot 5, 1:1,000), anti-demPP2A (Merck Millipore, clone 4b7, 05-577, lot 3154938, 1:500), anti-PP2A subunit C (Cell Signaling, clone 52F8, 2259T, lot 2 1:1,000) and anti-PME-1 (Merck Millipore, 07-095, lot 2805155, 1:1,000). For analysis on the Jess system from Protein-Simple, anti-caspase-1 (Adipogen, clone Casper-1, AG-20B-0042, lot A28881708, 1:50) was used. For treatment of primary neuronal cultures, IL-1β receptor antagonist (R&D Systems), IRAK4 inhibitor PF06650833 (Sigma) and MEK inhibitor UO126 (Cell Signaling) were used.

Behavioural phenotyping

The Morris water maze test was conducted as previously described in detail³. Probe trial data at day 9 are represented as time spent in quadrant Q1, the quadrant in which the platform had been located. This time was compared with the average time spent in all other quadrants.

Biochemical extraction of mouse tissue

For detection of caspase-1 in cortex sample of mice, RIPA fractions were used. Brain pieces were homogenized in homogenization buffer (1× PBS, 5 mM NaF, 20 mM pyrophosphate) before addition of equal volumes of 2× RIPA buffer (50 mM Tris-HCl, 150 mM NaCl, 2% NP40, 1% sodium dodecylsulfate, 0.2% SDS). Samples were incubated on ice and centrifuged for 20 min at 21,000g before supernatants were transferred to a new tube and stored at -20 °C. Protein levels were determined by using a BCA assay (Thermo Fisher Scientific) according to the manufacturer's instructions. For the analysis of tau and kinases/phosphatase levels in the mouse brain, hippocampi were homogenized in 10 volumes H buffer (10 mM Tris-HCl, 1mM EGTA, 800 mM NaCl, 10% sucrose, 0.1 mM PMSF, 1mM sodium orthovanadate, 1× protease/phosphatase inhibitor (Cell Signaling)) by using a Precellys device (Bertin Instruments). The homogenate was centrifuged for 20 min at 21,000g and the supernatant collected. The pellet was extracted again in the same volume of H buffer, centrifuged and the supernatant combined with the previous one. The pellet was homogenized in RIPA buffer for the analysis of membrane-bound proteins. The combined supernatant was supplemented with 1% sarkosyl and incubated for 2 h at 37 °C. Samples were centrifuged for 1 h at 300,000g and the supernatant, which contained sarkosyl-soluble tau, collected, aliquoted and stored at -80 °C.

ELISA quantification of pro-inflammatory cytokines

The protein from the snap-frozen tissue was extracted in RIPA buffer as described above. A BCA (Thermo Fisher Scientific) was performed on the supernatants containing the extracted proteins from the peripheral organs and these were equalized to the lowest sample within each organ group, namely spleen (1.3 mg ml⁻¹), medial intestine (1.2 mg ml⁻¹) and colon (1 mg ml⁻¹). The levels of IL-1β, TNF and IL-6 were determined on these and the serum using an electrochemoluminescence ELISA (Meso Scale Discovery). As per the manufacturer's instructions, samples were diluted 1:1 onto the plate using reagent diluent 41 supplied with the kit.

Human tissue samples

Post-mortem brain samples from patients with confirmed FTD as well as age-matched controls provided by the Biobank of Hospital Clínic, Barcelona, IDIBAPS were used according to their guidelines. All experiments with human samples complied with relevant ethical regulations. An informed consent was signed by all patients and they had agreed to the use of their brain material for medical research. Frozen brain pieces were either sectioned for staining or homogenized and lysed in RIPA buffer as described for the mouse tissue.

Immunoblot analysis

For standard immunoblot analysis, samples were supplemented with 1× NuPAGE sample buffer (Thermo Fisher Scientific), heated for 5 min at 95 °C and loaded on 4–12% NuPAGE Novex gels (Thermo Fisher Scientific). After transfer of proteins to nitrocellulose membranes, membranes were blocked with 3% BSA in TBS followed by incubation with primary antibodies in 3% BSA in TBS-Tween. Visualization of proteins was achieved by using fluorescent-tagged secondary antibodies (LI-COR). Imaging was performed by using a LI-COR ODYSSEY CLx. Data were analysed with ImageStudio software version 5.2.5 (LI-COR). All controls were run as loading controls on the same gel.

Immunohistochemical staining

Mouse brain sections of 40- μm thickness were stained in a free-floating format. For each experiment, at least three serial sections per mouse were used. For 3,3'-diaminobenzidine (DAB) staining, sections were washed, incubated with 0.3% H_2O_2 before blocking in 3% milk in 1 \times TBS with 0.25% Triton-X100 (TBS-X). Sections were incubated with a biotinylated primary antibody in blocking buffer overnight at 4 °C. After washing, the ABC kit (VectorLabs) was applied according to the manufacturer's instructions. Three more wash steps followed before the sections were developed in DAB solution (VectorLabs) and then washed three more times. Sections were mounted on glass slides, dried overnight and dehydrated in an Ethanol series followed by Xylo. Cytoseal 60 (Thermo Fisher Scientific) was used to apply cover glasses. Imaging was done with a 10 \times objective on an AxioScan. Z1 slide scanner from Zeiss. Threshold-analysis was performed with ImageJ/Fiji software version 2.0.0.-rc-67/1.52c. For fluorescent stainings of mouse sections, sections were stained either free-floating or after mounting on slides. Sections were washed and pre-treated with 1 mg ml^{-1} pepsin in 0.2M HCl for 10 min at 37 °C whenever AT8 staining was applied. Blocking was performed with 3% BSA in 1 \times PBS and 0.1% Triton-X100 (PBS-T). Sections were incubated with primary antibodies in blocking buffer overnight at 4 °C. After washing, AlexaFluor-labelled secondary antibodies (Invitrogen) in blocking solution were applied, followed by additional washing steps. To quench autofluorescence, the sections were incubated for 20 min in 0.1% Sudan Black B (Sigma) in 70% ethanol and washed extensively in 1 \times PBS. For thioflavin-S staining, the quenching was followed by a 5 min incubation in 0.025% thioflavin S (Sigma) in 50% ethanol, quick washes in 50% ethanol and a longer wash step in water. Prolong Gold and DAPI (Invitrogen) was used as a mounting medium. For human frozen sections, epitope retrieval was achieved by heating the sections in 10 mM citrate buffer, pH 6. Sections were blocked with 10% donkey serum in PBS-T, followed by incubation with primary antibodies in 1% donkey serum in PBS-T overnight at 4 °C. After washing, AlexaFluor-labelled secondary antibodies (Invitrogen) in 1% donkey serum in PBS-T were added, followed by incubation with Hoechst and TrueBlack Lipofuscin Autofluorescence Quencher (Biotium). After extensive washing, sections were coverslipped with Immu-Mount (Thermo Fisher Scientific).

When full hippocampus pictures were necessary, images were taken with a Zeiss Axio Scan.Z1 with a 20 \times objective or with a Nikon eclipse Ti with a 20 \times objective. For high-magnification images, a Zeiss LSM700 with a 40 \times or 60 \times oil objective was used. Image processing was achieved by using ImageJ Version 2.0.0.-rc-67/1.52c, Adobe Photoshop CS5 Version 12.0.1 and Imaris Version 9.

PCR on faecal samples

The DNA was extracted from the faecal samples obtained from the colon using the QIAamp PowerFecal DNA Kit (Qiagen) following the manufacturer's instructions. The DNA was quantified spectrophotometrically and equalized to 10 ng μl^{-1} . The PCR was performed with 4 μl of DNA, 2.5 μl of forward and reverse primers, 12.5 μl of Kappa Sybr Fast mix (Thermo Fisher Scientific), 0.5 μl Rox High and 4.5 μl ultrapure water per sample. A standard curve was also prepared using known concentrations of *Escherichia coli*. The PCR was carried out on a StepOne Plus real-time PCR system (Applied Biosystems) under the following conditions: 95 °C for 3 min, and 60 cycles of 95 °C for 3 s, 64 °C for 30 s. The following probes were used to determine the V6 region of bacterial 16S, forward primers: 5'-CNACGCGAAGAACCTTANC-3', 5'-ATACGCGARGAACCTTACC-3', 5'-CTAACCGANGAACCTYACC-3', 5'-CAACGCGMARAACTTACC-3', and reverse primer: 5'-CGACRRCCATGCANCACT-3' (Metabion International AG). The analysis was performed using StepOne software version 2.1 (Applied Biosystems).

Primary neuronal cultures

Primary mouse neuronal cultures were prepared from newborn pups (P0) from C57BL/6N or Tau22 mice. Brains were taken out, meninges removed and hippocampi dissected out. Hippocampi were washed in 1 \times HBSS (Thermo Fisher Scientific) and single-cell suspensions generated by incubation with trypsin and DNase before careful trituration. Seventy thousand cells per well were plated on poly-D-lysine-coated 24-well plates in Neurobasal medium (Thermo Fisher Scientific) supplemented with B-27 (Thermo Fisher Scientific) and used for experiments at DIV12-14.

Primary microglia cultures

Primary mouse microglia cell cultures were prepared as previously described³ from wild-type, *Asc^{-/-}* and *Nlrp3^{-/-}* pups. In brief, mixed glia cultures were prepared from newborn mice (P0-P4). Cells were plated in DMEM (Thermo Fisher Scientific) with 10% FCS, 10% L929 conditioned medium and 100 U ml^{-1} penicillin-streptomycin (Thermo Fisher Scientific). Seven to ten days after cultivation, microglia were collected by shake off, counted and plated in DMEM supplemented with 1% N2. If conditioned medium was used to treat neurons, microglia medium was replaced with Neurobasal medium (Thermo Fisher Scientific) supplemented with B-27 (Thermo Fisher Scientific) at least 12 h before applying treatments. Microglia were shaken off up to three times.

Recombinant tau preparation

Human wild-type (2N4R) or P301S (1N4R) tau-expressing, inducible plasmids were transformed into BL21(DE3) *E. coli* (Agilent). Single colonies were grown to a suitable density and tau expression induced by addition of IPTG (Merck). After 3 h, bacteria were pelleted and resuspended in BRB-80 (80 mM PIPES, 1 mM magnesium sulfate, 1 mM EGTA, pH 6.8) supplemented with 0.1% 2-mercapthoethanol and 1 mM PMSF. The resuspension was sonicated, followed by a centrifugation to remove the bacterial debris. The tau-containing supernatant was boiled for 10 min and centrifuged again. Supernatants were applied to a cation exchange chromatography column, which was washed afterwards with BRB-80 supplemented with 0.1% 2-mercapthoethanol. Tau was eluted with increasing concentrations of sodium chloride dissolved in BRB-80 with 0.1% 2-mercapthoethanol and each fraction of the elution tested for the presence of tau by a Coomassie gel stain. Tau-containing fractions were combined and the buffer replaced with 10 mM ammonium bicarbonate by using Amicon ultra centrifugal units (10-kDa molecular weight cut-off). Buffer replacement was performed several times to efficiently remove all remaining endotoxins. The last concentrated solution was collected from the columns, protein concentration assessed by a BCA assay (Thermo Fisher Scientific), aliquoted and dried in a speed vac before storage at -80 °C. Tau proteins were resuspended in DPBS right before use and each preparation was tested for endotoxin levels with an endotoxin quantification kit (Pierce) following the manufacturer's instructions. Tau oligomers were prepared as previously described²⁶.

Tau and A β fibril formation

Recombinant human tau was fibrillized as previously described²⁷. In brief, 8 μM human tau was incubated with 10 mM DTT, 10 mM HEPES, 100 mM NaCl and 8 μM heparin (1:1 ratio to tau) for 72 h at 37 °C. The solution was centrifuged at 100,000g for 1 h and the pellet resuspended in DPBS. To generate A β fibrils, HFIP-treated A β_{1-42} (Bachem) was dissolved in sterile PBS to a final concentration of 250 μM and incubated on a shaker for 84 h at 37 °C. All fibril preparations were aliquoted and frozen at -80 °C for single use.

Treatment of microglia

For IL-1 β analysis, 75,000 microglia per well in 96-well plate were used. For caspase-1 detection or collection of conditioned medium

Article

for neuronal experiments, 2 million microglia or 600,000 microglia per well were plated in a 6-well plate. Microglia were activated with 100 ng ml⁻¹ ultrapure LPS (*E. coli* 0111:B4, Invivogen) for 3 or 3.5 h. After washing with warm DPBS, recombinant tau monomers, oligomers or fibrils at a concentration of 2 μ M or 10 μ M A β fibrils were added for 6 h to the cultures or 10 mM ATP for 30 minutes before collection of conditioned media. For CRID3 (Sigma Aldrich) treatments, the inhibitor was mixed in with the tau treatments at a concentration of 100 nM. Conditioned media were analysed for IL-1 β secretion by ELISA (R&D Systems) according to the manufacturer's instructions or for caspase-1 secretion by using a Jess system (ProteinSimple). Here, samples were run undiluted on a 12–230-kDa separation module according to the manufacturer's instructions and detection was achieved by using a caspase-1 antibody (Adipogen) at 1:50. Data were analysed with Compass software version 4.0.0 (ProteinSimple).

Stereotaxic surgery

Intracerebral stereotaxic injections of brain material were performed as previously described³. In brief, 3-month-old mice were anaesthetized with ketamine and xylazine before two small holes were drilled into the skull using a Dremel device attached to a stereotaxic frame. Mice received a bilateral injection of 2 μ l brain homogenate from APP/PS1 or wild-type mice (AP –2.5 mm, L \pm 2 mm, DV –1.8 mm; injected with 0.5 μ l min⁻¹). After suturing, mice were monitored until full recovery from the anaesthesia and housed under standardized conditions for 5 months before being killed.

RNA expression analysis through NanoString analysis

To isolate RNA, QIAzol Lysis Reagent (Qiagen) was added to brain homogenates of mice at different ages. After the addition of chloroform, samples were shaken vigorously and centrifuged for 15 min at 12,000g. The aqueous phase was collected and the RNA isolated with an RNeasy Mini kit (Qiagen) according to the manufacturer's instructions. One hundred nanograms of total RNA was hybridized using the NanoString Mouse Neuroinflammation V1.0 panel for 24 h following the manufacturer's instructions. The hybridized RNA was then bound to a cartridge, analysed on an nCounter and the fields of view (FOV) count was set to 555.

Bioinformatic analysis of NanoString data

RCC files were imported into nSolver Analysis software 4.0 and processed by the Advanced Analysis 2.0.115 software package. geNORM identified the ideal housekeepers *Mto1*, *Cnot10*, *Lars*, *Ccdc127*, *Fam104a*, *Tada2b* and *Tbp*. Other housekeeping genes were discarded because of high variances in gene expression throughout all samples. The geometric mean of negative controls was subtracted from all counts. The geometric mean of positive controls and housekeepers were used to normalize the data. Transcripts with an expression value of ≤ 20 were excluded from the analysis. Expression profiles of 578 transcripts in wild-type and Tau22 samples were used to generate a network based on co-expression consisting in Biolayout express 3D v3.3. Fold changes in gene expression patterns were visualized in form of colour-coded nodes in Cytoscape v3.6.1. SOM clustering of expression profiles of 671 genes into a matrix consisting of 10 \times 10 clusters and ANOVA were performed with Partek Genomics Suite v6.6 and 7.18. Genes of clusters with elevated expression levels were grouped and defined as gene signatures

for Tau22 mice at 3 and 11 months of age. Brain related interferon regulated genes were identified via the interferome data base v2.01. Gene ontology (GO) analysis for gene signatures were performed in STRING v11.0. Boxplots and gene rank plots were generated in Excel.

Statistics and reproducibility

Each *n* represents an independent biological sample. All graphs show means \pm s.e.m. Statistical analysis was performed with GraphPad Prism software version 7.0c, applying either a two-tailed unpaired *t*-test or one-way or two-way ANOVA with Tukey's post hoc test comparing all groups to each other. The GO enrichment analysis was performed according to Fisher's exact test followed by a correction for multiple testing²⁸.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated and/or analysed during this study are either included in this article (and its Supplementary Information) or are available from the corresponding author on reasonable request. Source Data for Figs. 1–4 and Extended Data Figs. 1, 5–10 are provided with the paper.

25. Kanneganti, T.-D. et al. Bacterial RNA and small antiviral compounds activate c through cryopyrin/Nalp3. *Nature* **440**, 233–236 (2006).
26. Lasagna-Reeves, C. A., Castillo-Carranza, D. L., Guerrero-Muoz, M. J., Jackson, G. R. & Kaye, R. Preparation and characterization of neurotoxic tau oligomers. *Biochemistry* **49**, 10039–10041 (2010).
27. Ising, C. et al. AAV-mediated expression of anti-tau scFvs decreases tau accumulation in a mouse model of tauopathy. *J. Exp. Med.* **214**, 1227–1238 (2017).
28. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45** (D1), D362–D368 (2017).

Acknowledgements This work was supported by funding from the Deutsche Forschungsgemeinschaft (DFG) to C.I. (IS 299/3-1) and under Germany's Excellence Strategy – EXC2151 – 390873048. R.K. received funding from a NIH grant (R01 AG054025), and D.G. and M.T.H. received further funding from a NIH grant (R01 AG059752-02). We thank I. Rácz for help with obtaining approval by the local ethical committee for the animal experiments; P. Davies for providing the MC1 and PHF-1 antibodies; the DZNE light microscope facility (LMF) for providing microscopes and advice; and the DZNE Image and Data Analysis Facility (IDAF) for providing analysis computers, software and advice.

Author contributions C.I. and M.T.H. designed most of the experiments; C.I., C.V., S.Z. and H.S. performed experiments and analysed data with assistance of A.V.-S., A.G., F.S. and M.M.; D.T. quantified microglia morphology and performed ASC speck experiments; S.A. performed and analysed microglia treatments with tau; R.M.M. performed and analysed microbiome experiments; S.S. performed behaviour experiments; F.B. validated antibodies and helped with IL-1 β analyses; S.O. provided neuron cultures; S.V.S. and J.S. performed analysis of the NanoString data; M.T.H. quantified ASC specks and analysed data; R.K. provided tau oligomers; D.B., D.T.G., E.L. and L.B. provided mice, samples and advice; C.I. and M.T.H. wrote the manuscript with input from all co-authors.

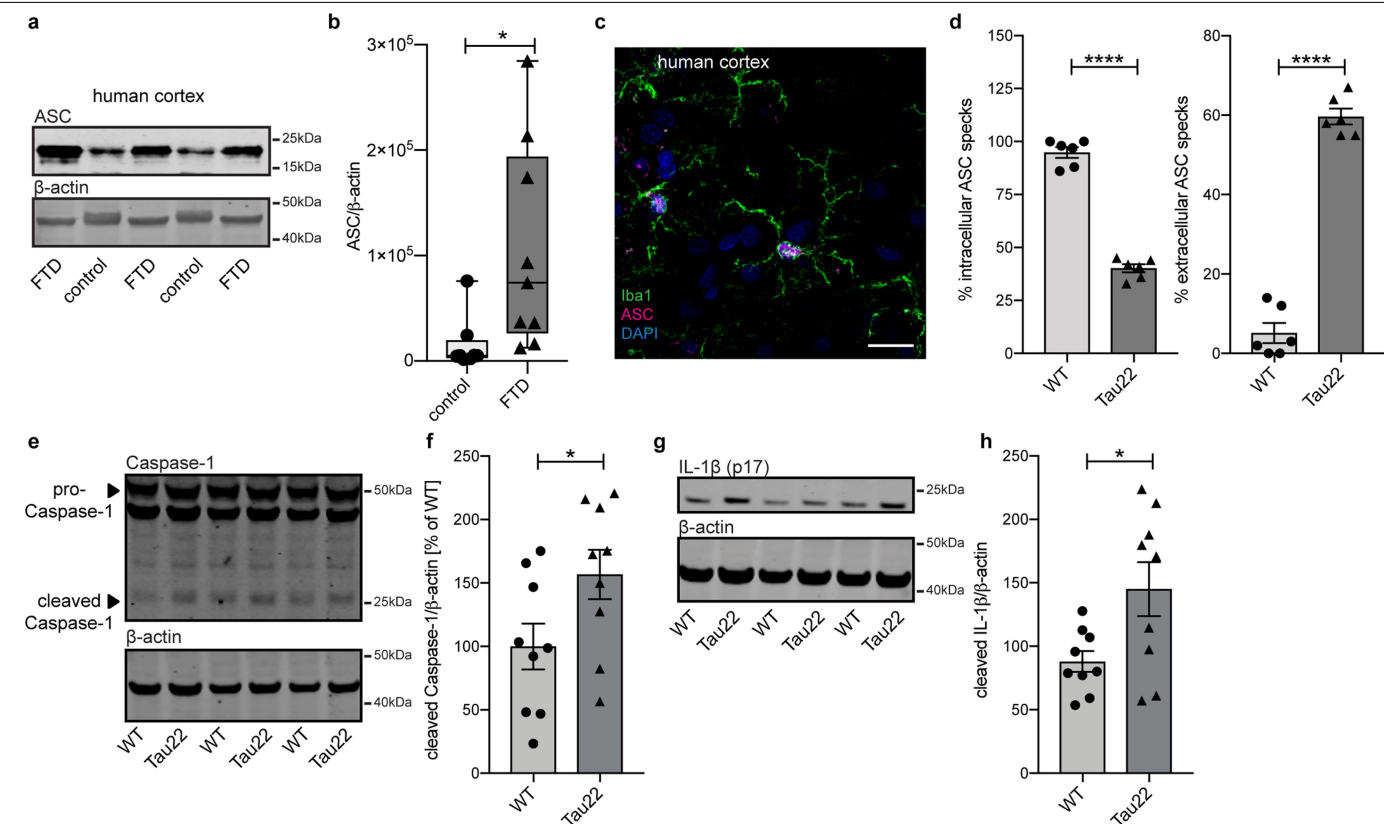
Competing interests E.L. is a co-founder and advisor, J.S. is an employee and M.T.H. serves as an advisory board member at IFM Therapeutics. M.T.H. is an advisory board member at Alektor. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1769-z>.

Correspondence and requests for materials should be addressed to M.T.H.

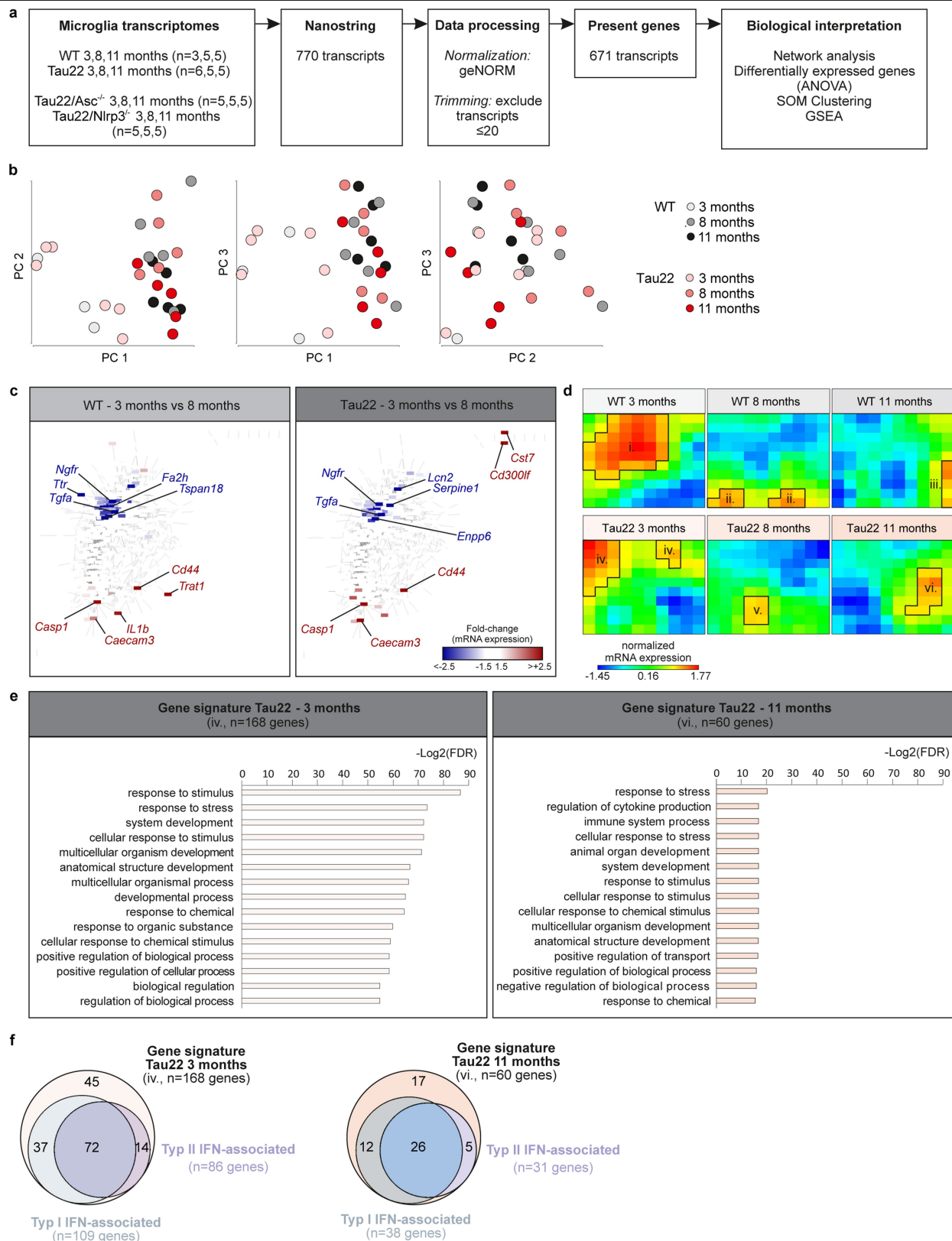
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | The NLRP3 inflammasome is activated in Tau22 mice.

a, Immunoblot analysis of ASC and β -actin in human cortex of patients with FTD and control patients. **b**, Quantification of data from **a**. Box plots show 25th and 75th percentiles. $n = 8$ for controls, $n = 9$ for FTD, $*P = 0.0239$. **c**, Immunohistochemical staining of human cortex from a patient with FTD for microglia and ASC. $n = 3$. Scale bar, 20 μ m. **d**, Quantification of percentage of intracellular ASC and of extracellular ASC specks from staining shown in Fig. 1g. $n = 6$ mice per group, $****P < 0.0001$. **e**, Immunoblot analysis of hippocampus

samples of 8-month-old wild-type and Tau22 mice stained for caspase-1 and β -actin. **f**, Quantification of data from **e**. $n = 9$ per group, $*P = 0.0489$. **g**, Immunoblot analysis of hippocampus samples of 8-month-old wild-type and Tau22 mice stained for IL-1 β (p17) and β -actin. **h**, Quantification of data from **g**. $n = 9$ per group, $*P = 0.0236$. For gel source data, see Supplementary Fig. 1. All graphs are presented as mean \pm s.e.m. and were analysed by two-tailed unpaired t -test.



Extended Data Fig. 2 | Gene signatures in wild-type and Tau22 mice identified by NanoString analysis. **a**, Workflow for NanoString analysis. **b**, Two-dimensional principal component (PC) analysis of wild-type and Tau22 mice at 3, 8 and 11 months of age. **c**, Gene network analysis of regulated genes at 3 versus 8 months in wild-type and Tau22 mice identified by NanoString

analysis. **d**, SOM clustering of wild-type and Tau22 mice at 3, 8 and 11 months of age with definition of gene sequences (i)–(vi). **e**, Gene signatures in 3-month-old Tau22 mice defined by cluster (iv) and in 11-month-old Tau22 mice defined by cluster (vi). Fisher's exact test followed by a correction for multiple testing. **f**, Interferome Venn diagrams based on cluster (iv) and (vi) in Tau22 mice.

a

Legend for color coded edges:

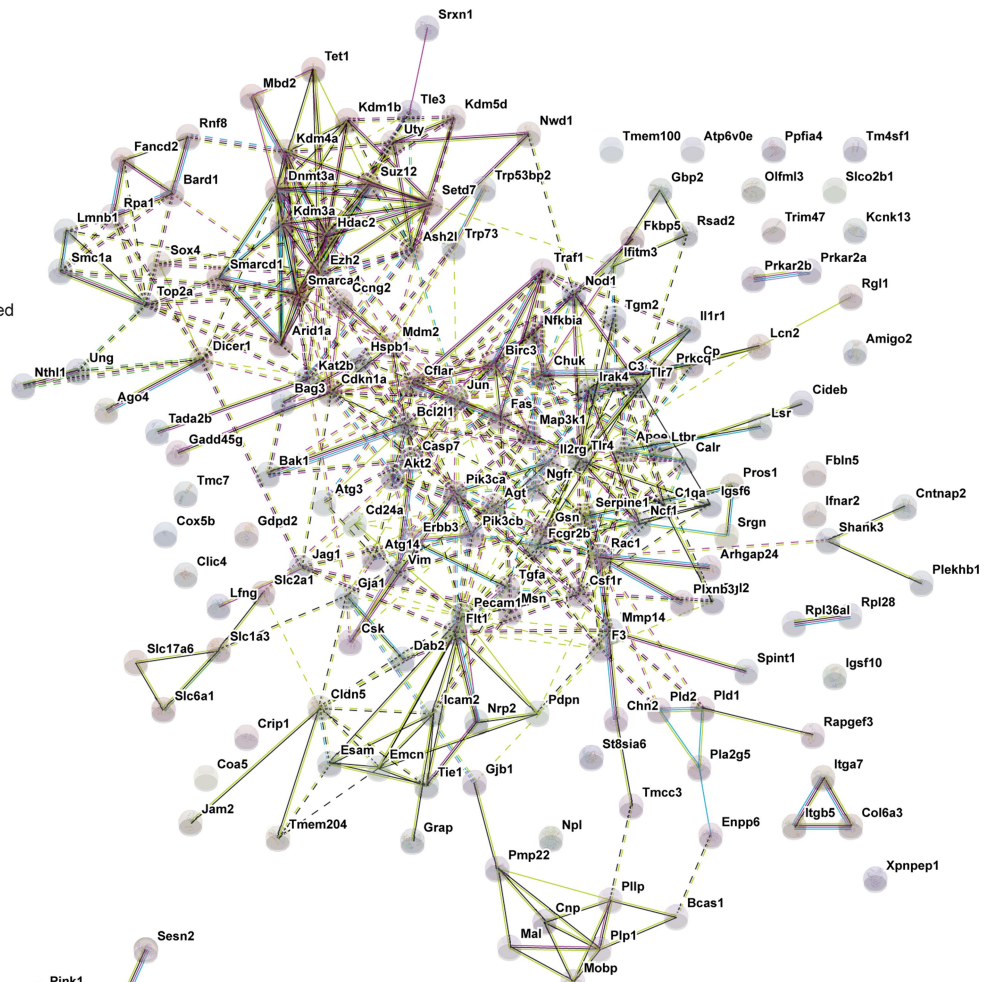
- text mining
- co-expression
- protein homology

Predicted interactions:

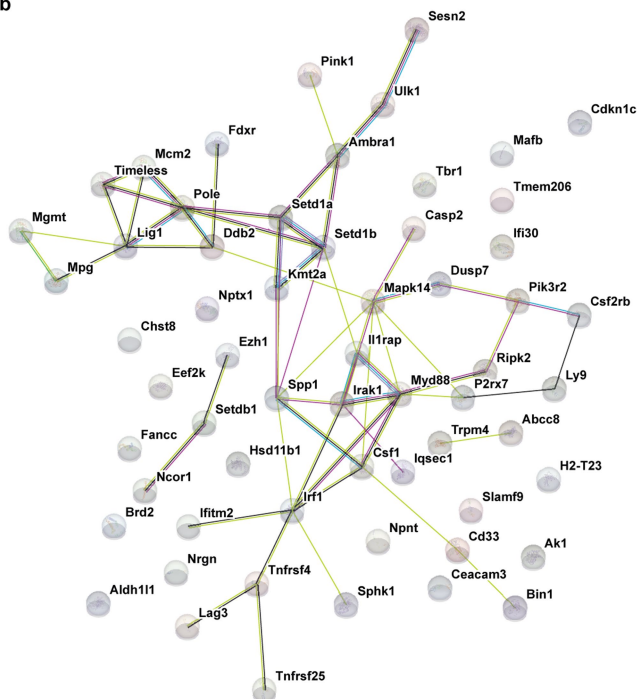
- gene neighborhood
- gene fusions
- gene co-occurrence

Known interactions:

- from curated databases
- experimentally determined

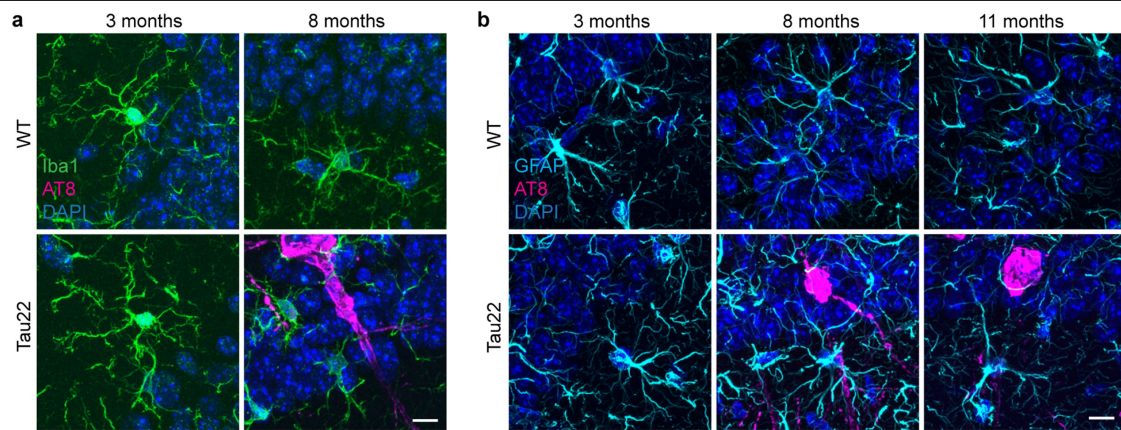


b



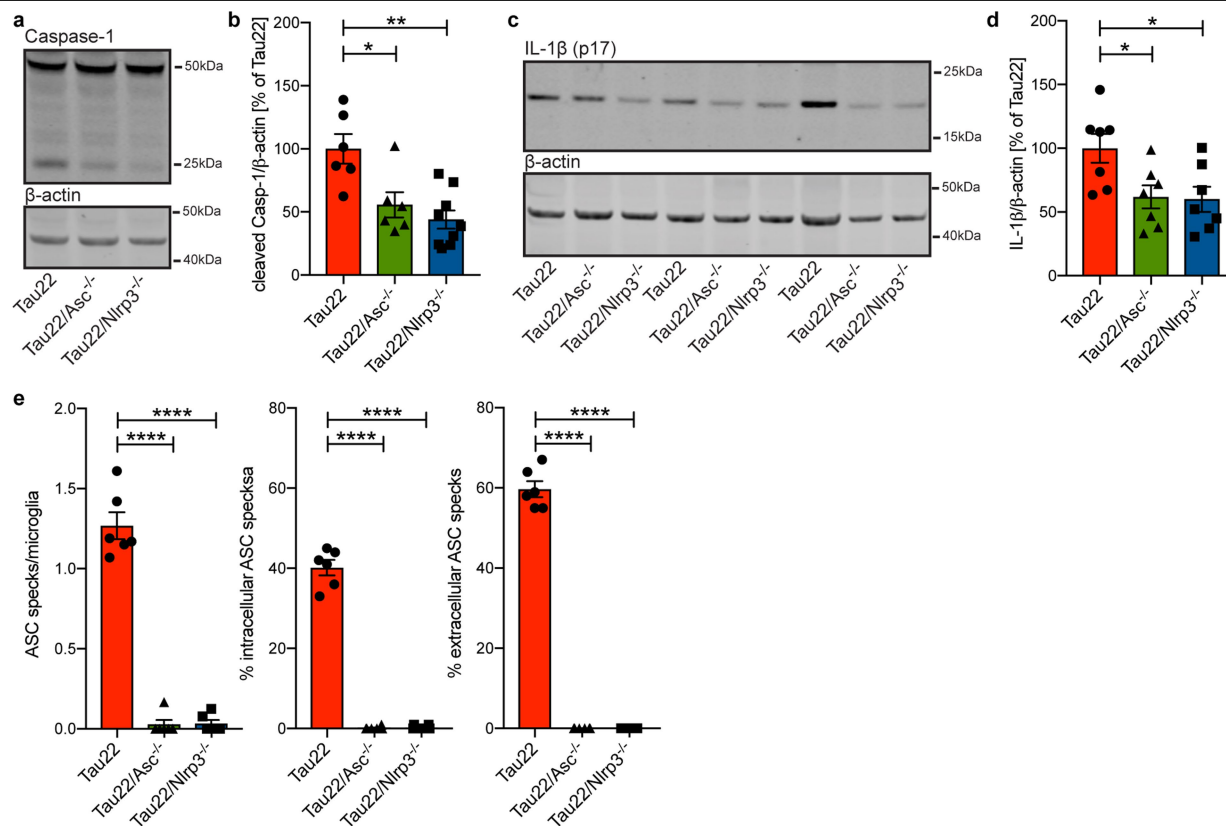
Extended Data Fig. 3 | STRING network analysis of Tau22 mice. a, b, Networks visualizing the functional protein association for gene signatures in Tau22 mice at 3 (a) and 11 (b) months of age. Nodes in the network represent proteins.

Edges represent protein–protein interactions, which (depending on the colour) indicate known or predicted interactions.



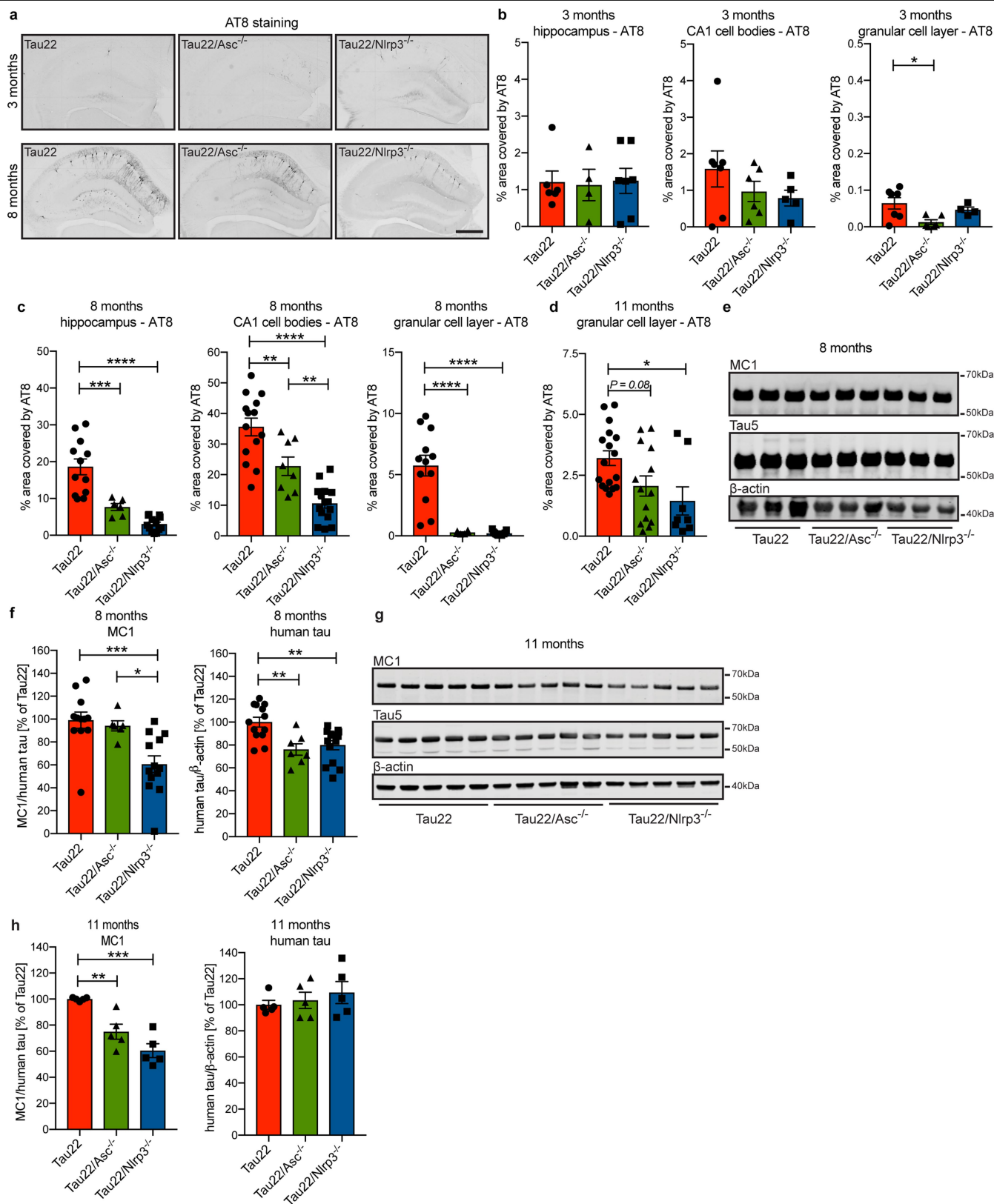
Extended Data Fig. 4 | Astrocyte morphology does not change in Tau22 mice. a, Immunohistochemical staining for microglia (IBA1) and phosphorylated tau (AT8) in wild-type and Tau22 mice at 3 and 8 months of age. *n* = 8. Scale bar, 10 μ m.

b, Immunohistochemical staining for astrocytes (GFAP) and phosphorylated tau (AT8) in wild-type and Tau22 mice at 3, 8 and 11 months of age. *n* = 8. Scale bar, 10 μ m.



Extended Data Fig. 5 | Knockout of *Asc* or *Nlrp3* efficiently inhibits NLRP3 inflammasome function. **a**, Immunoblot analysis of hippocampus samples from 11-month-old mice stained for caspase-1 and β-actin. **b**, Quantification of data from **a**. $n = 6$ for Tau22 and Tau22/Asc^{-/-}, $n = 9$ for Tau22/Nlrp3^{-/-}, * $P = 0.0156$, ** $P = 0.0012$. **c**, Immunoblot analysis of hippocampus samples of 11-month-old mice stained for IL-1β (p17) and β-actin. **d**, Quantification of data from **c**. $n = 7$ for all groups. Tau22 versus Tau22/Asc^{-/-}: * $P = 0.0399$, Tau22 versus

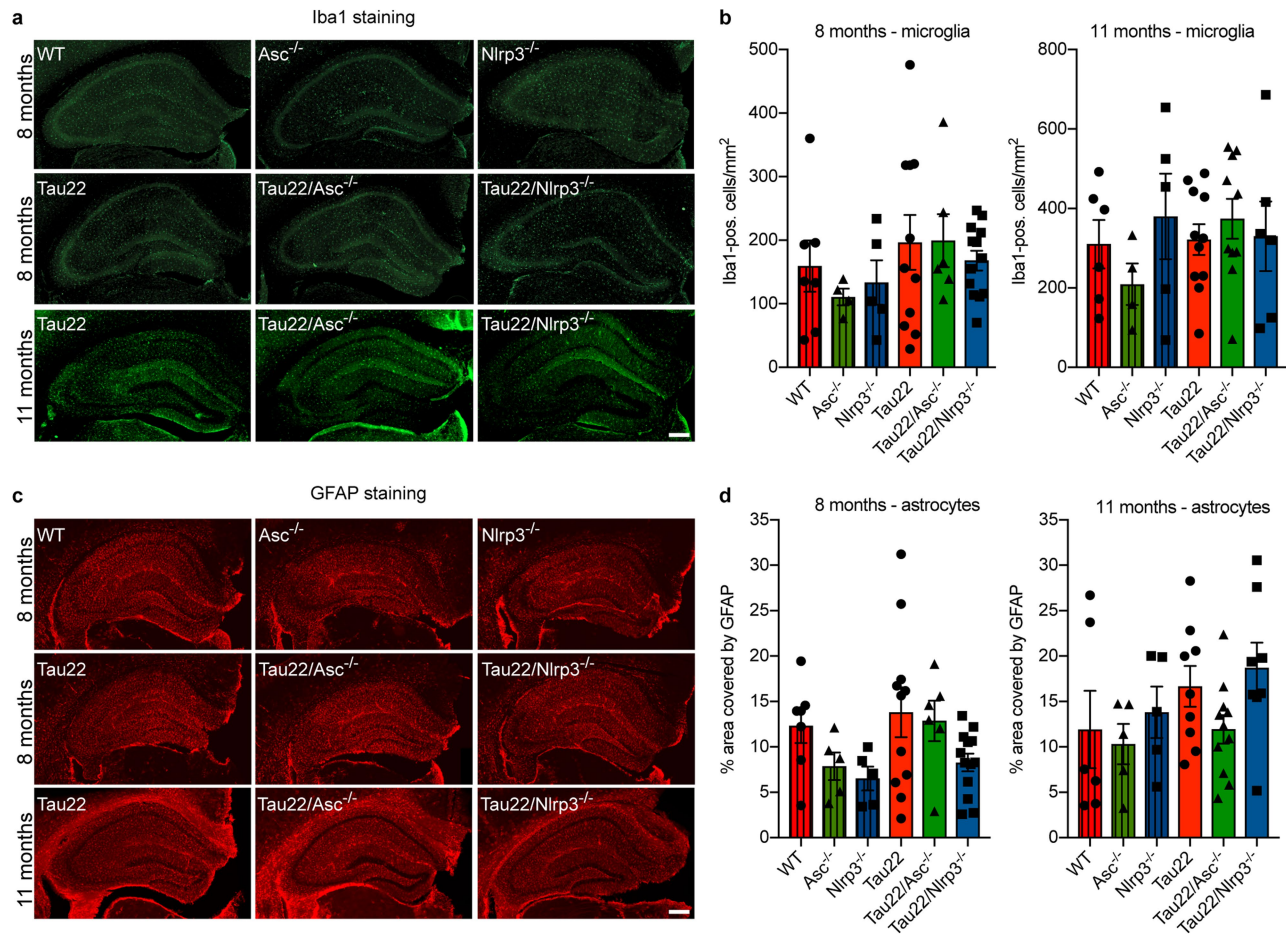
Tau22/Nlrp3^{-/-}: * $P = 0.0310$. **e**, Quantification of number of ASC specks/microglia, percentage of intracellular ASC specks and percentage of extracellular ASC specks in hippocampus sections of 11-month-old mice. $n = 6$ mice per group, **** $P < 0.0001$ for all comparisons. For gel source data, see Supplementary Fig. 1. All graphs are presented as mean ± s.e.m. and were analysed by one-way ANOVA followed by Tukey's test.



Extended Data Fig. 6 | See next page for caption.

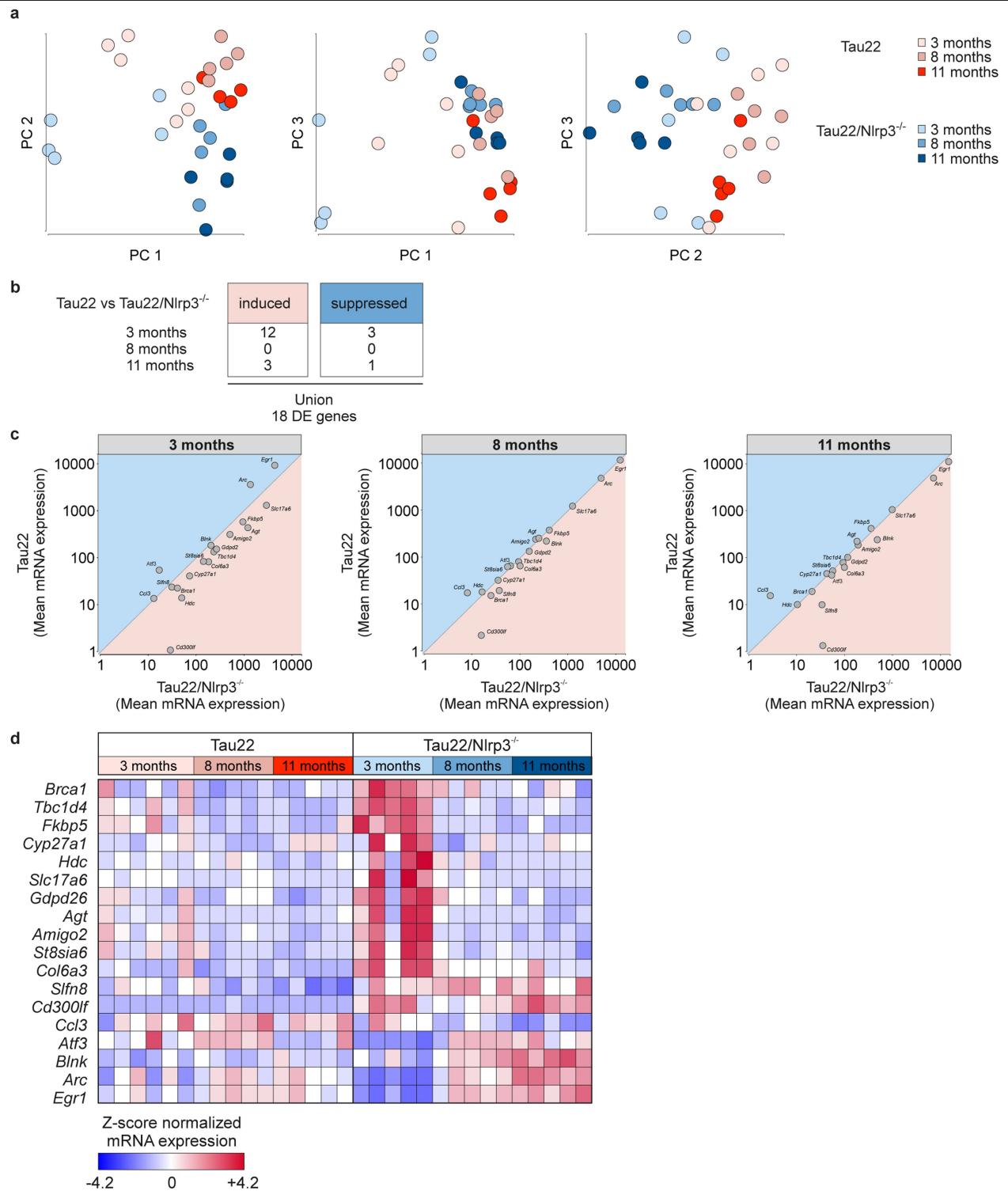
Extended Data Fig. 6 | Tau pathology is reduced in inflammasome-knockout mice. **a**, Immunohistochemical staining for phosphorylated tau (AT8) in mouse hippocampi. Scale bar, 500 μ m. **b**, Quantification of AT8 in hippocampus, CA1 cell body layer and granular cell layer in the dentate gyrus of 3-month-old mice shown in **a** ($n = 6$ for hippocampus Tau22 and CA1 and dentate gyrus Tau22/*Asc*^{-/-}, $n = 4$ for hippocampus Tau22/*Asc*^{-/-} and dentate gyrus Tau22/*Nlrp3*^{-/-}, $n = 7$ for hippocampus Tau22/*Nlrp3*^{-/-} and CA1 and dentate gyrus Tau22, $n = 5$ for CA1 Tau22/*Nlrp3*^{-/-}). * $P = 0.0181$. **c**, Quantification of AT8 in hippocampus, CA1 cell body layer and granular cell layer in the dentate gyrus of 8-month-old mice shown in **a**. Hippocampus: $n = 12$ for Tau22, $n = 6$ for Tau22/*Asc*^{-/-}, $n = 13$ for Tau22/*Nlrp3*^{-/-}. *** $P = 0.0004$ and **** $P < 0.0001$. CA1: $n = 14$ for Tau22, $n = 8$ for Tau22/*Asc*^{-/-}, $n = 15$ for Tau22/*Nlrp3*^{-/-}. Tau22 versus Tau22/*Asc*^{-/-}: ** $P = 0.0052$, Tau22 versus Tau22/*Nlrp3*^{-/-}: **** $P < 0.0001$, Tau22/*Asc*^{-/-} versus Tau22/*Nlrp3*^{-/-}: ** $P = 0.0075$. Dentate gyrus: $n = 12$ for Tau22, $n = 6$ for Tau22/*Asc*^{-/-}, $n = 13$ for Tau22/*Nlrp3*^{-/-}, **** $P < 0.0001$. **d**, Quantification

of AT8 in granular cell layer in the dentate gyrus of 11-month-old mice shown in Fig. 2a. $n = 17$ for Tau22, $n = 14$ for Tau22/*Asc*^{-/-}, $n = 8$ for Tau22/*Nlrp3*^{-/-}, * $P = 0.0196$. **e**, Immunoblot analysis of sarkosyl-soluble fraction of hippocampi from 8-month-old Tau22, Tau22/*Asc*^{-/-} and Tau22/*Nlrp3*^{-/-} mice stained for misfolded tau (MC1), total tau (Tau5) and β -actin. **f**, Quantification of data from **e**. MC1: $n = 12$ for Tau22, $n = 6$ for Tau22/*Asc*^{-/-}, $n = 13$ for Tau22/*Nlrp3*^{-/-} with Tau22 versus Tau22/*Nlrp3*^{-/-}: *** $P = 0.0009$ and Tau22/*Asc*^{-/-} versus Tau22/*Nlrp3*^{-/-}: * $P = 0.0190$. Human tau: $n = 13$ for Tau22, $n = 7$ for Tau22/*Asc*^{-/-}, $n = 14$ for Tau22/*Nlrp3*^{-/-}. Tau22 versus Tau22/*Asc*^{-/-}: ** $P = 0.0047$, Tau22 versus Tau22/*Asc*^{-/-}: ** $P = 0.0037$. **g**, Immunoblot detection of misfolded tau (MC1), total tau (Tau5) and β -actin in sarkosyl-soluble fraction of mouse hippocampi at 11 months of age. **h**, Quantification of data from **g**. $n = 5$, ** $P = 0.0056$, *** $P = 0.0001$. For gel source data, see Supplementary Fig. 1. All graphs are presented as mean \pm s.e.m. and were analysed by one-way ANOVA followed by Tukey's test.



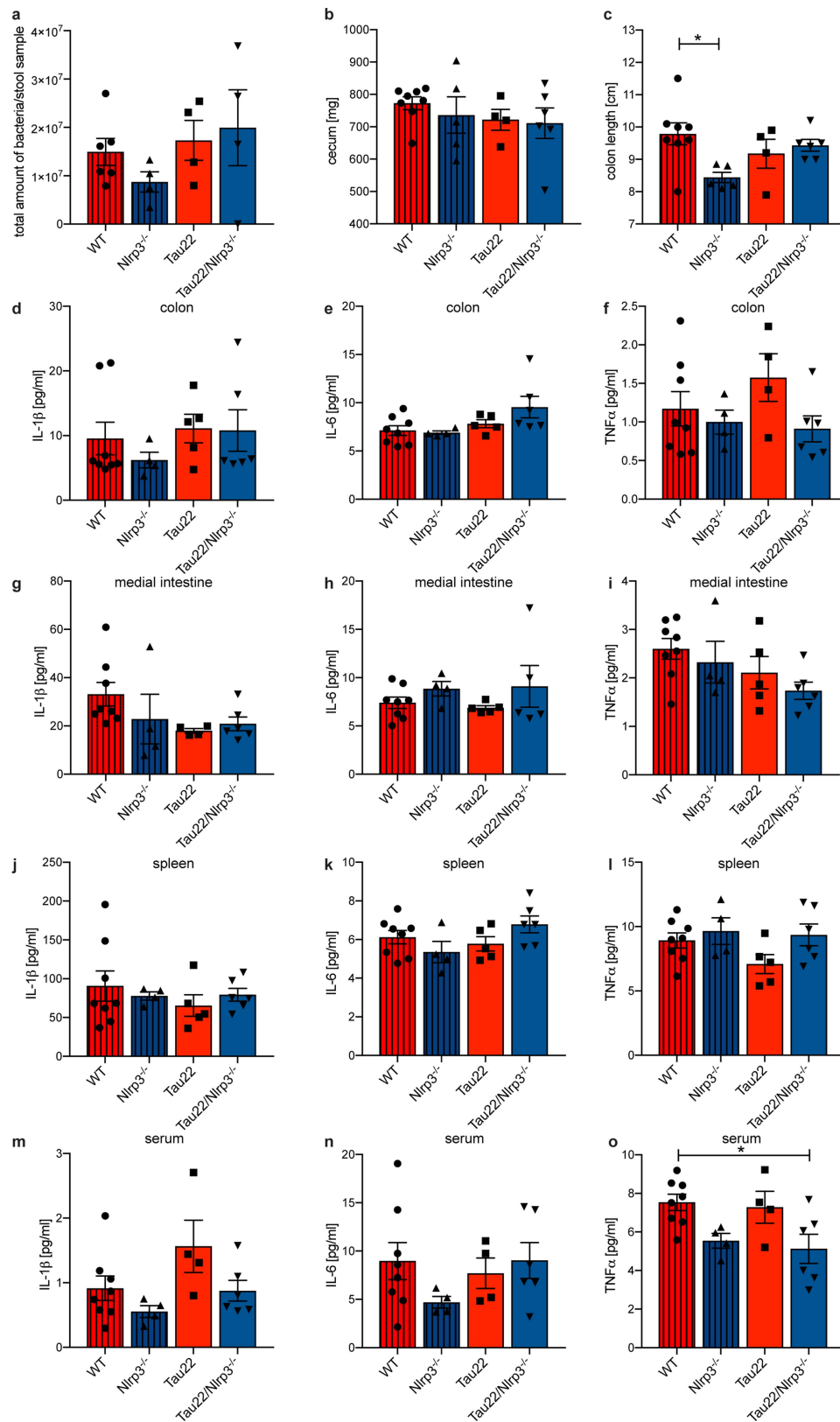
Extended Data Fig. 7 | Microglia and astrocyte numbers are unaltered in *Tau22/Asc*^{-/-} and *Tau22/Nlrp3*^{-/-} mice. **a**, Immunohistochemical staining of hippocampus of mice with the indicated genotypes and at the indicated ages for microglia (IBA1). Scale bar, 250 μ m. **b**, Quantification of IBA1-positive cells in the hippocampus as seen in **a** at 8 (left) and 11 months of age (right). $n = 7$ for 8 months wild-type, $n = 6$ for 11 months wild-type, $n = 4$ for 8 and 11 months *Asc*^{-/-}, $n = 5$ for 8 and 11 months *Nlrp3*^{-/-}, $n = 11$ for 8 and 11 months Tau22, $n = 6$ for 8 months Tau22/*Asc*^{-/-}, $n = 10$ for 11 months Tau22/*Asc*^{-/-}, $n = 13$ for 8 months Tau22/*Nlrp3*^{-/-}, $n = 6$ for 11 months Tau22/*Nlrp3*^{-/-}. **c**, Immunohistochemical

staining of hippocampus of mice with the indicated genotypes and at the indicated ages for astrocytes (GFAP). Scale bar, 250 μ m. **d**, Quantification of GFAP in the hippocampus as seen in **c** at 8 (left) and 11 months of age (right). $n = 7$ for 8 months wild-type, $n = 6$ for 11 months wild-type, $n = 5$ for 8 and 11 months *Asc*^{-/-} and *Nlrp3*^{-/-}, $n = 11$ for 8 months Tau22, $n = 9$ for 11 months Tau22, $n = 6$ for 8 months Tau22/*Asc*^{-/-}, $n = 11$ for 11 months Tau22/*Asc*^{-/-}, $n = 13$ for 8 months Tau22/*Nlrp3*^{-/-}, $n = 8$ for 11 months Tau22/*Nlrp3*^{-/-}. All graphs are presented as mean \pm s.e.m. and were analysed by one-way ANOVA followed by Tukey's test.



Extended Data Fig. 8 | Gene signatures in Tau22 and Tau22/*Nlrp3*^{-/-} mice identified by NanoString analysis. a, Two-dimensional PC analysis of Tau22 and Tau22/*Nlrp3*^{-/-} mice at 3, 8 and 11 months of age. *n* = 5 independent samples for each group. **b**, Number of induced or suppressed genes comparing Tau22

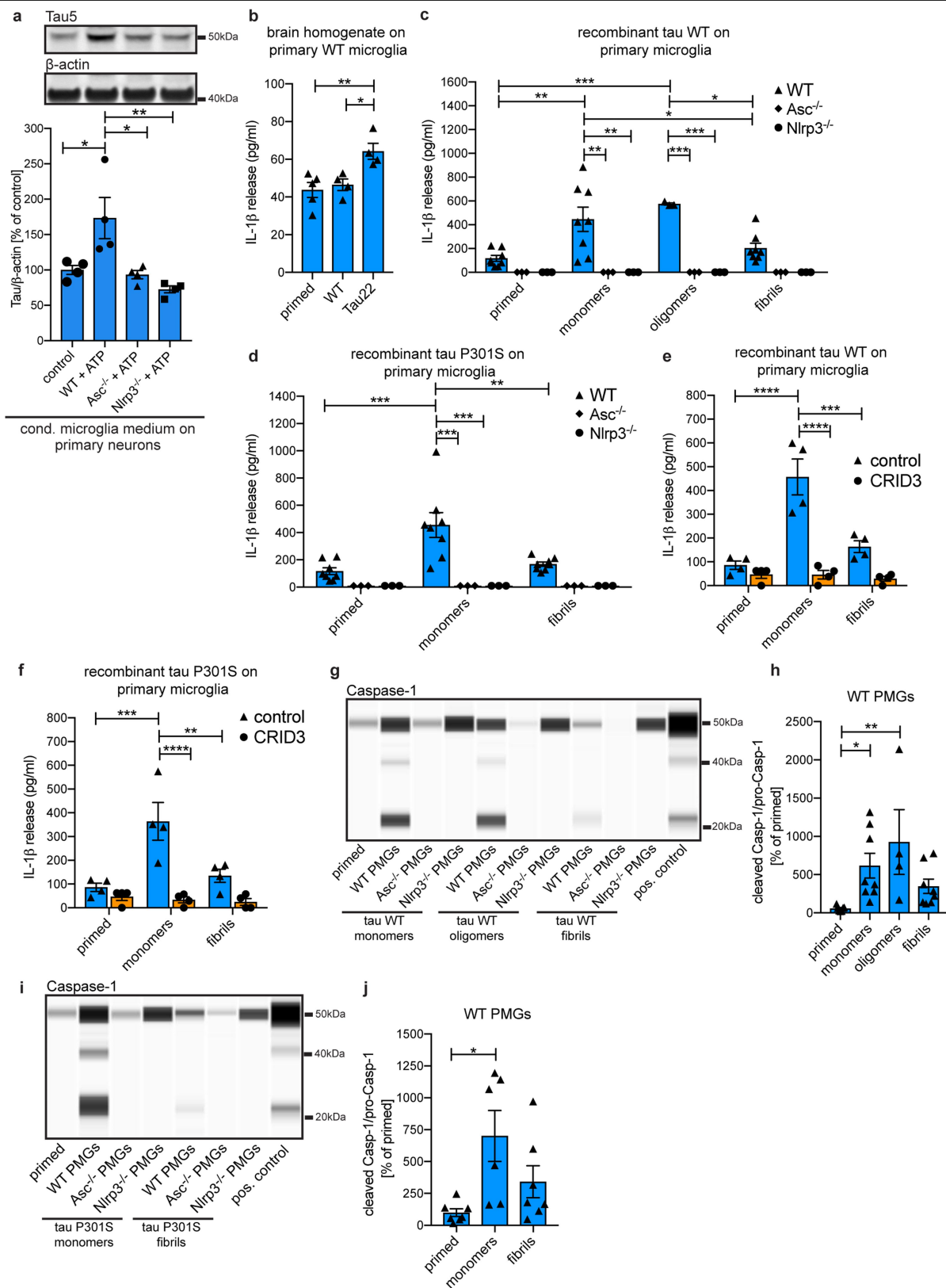
versus Tau22/*Nlrp3*^{-/-} at 3, 8 and 11 months. **c**, Gene plots of Tau22 versus Tau22/*Nlrp3*^{-/-} at 3, 8 and 11 months. **d**, Heat map comparing significantly changed genes in Tau22 versus Tau22/*Nlrp3*^{-/-} mice at various ages, identified by NanoString analysis.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | *Nlrp3*-knockout does not affect the microbiome of Tau22 mice. **a**, Amount of bacteria in stool samples obtained from the colon of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 6 for wild type, *n* = 4 for *Nlrp3*^{-/-}, Tau22, Tau22/*Nlrp3*^{-/-}. **b**, Caecum weight of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 5 for *Nlrp3*^{-/-}, *n* = 4 for Tau22, *n* = 6 for Tau22/*Nlrp3*^{-/-}. **c**, Colon length of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 5 for *Nlrp3*^{-/-}, *n* = 4 for Tau22, *n* = 6 for Tau22/*Nlrp3*^{-/-}, **P* = 0.0350. **d-f**, IL-1β, IL-6 and TNF levels in colon samples of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 4 for *Nlrp3*^{-/-}, *n* = 5 for IL-1β and IL-6 in Tau22, *n* = 4 for TNF in Tau22, *n* = 6 for

Tau22/*Nlrp3*^{-/-}. **g-i**, IL-1β, IL-6 and TNF levels in medial intestine samples of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 4 for *Nlrp3*^{-/-}, *n* = 4 for IL-1β in Tau22, *n* = 5 for IL-6 and TNF in Tau22, *n* = 6 for IL-1β and TNF in Tau22/*Nlrp3*^{-/-}, *n* = 5 for IL-6 in Tau22/*Nlrp3*^{-/-}. **j-l**, IL-1β, IL-6 and TNF levels in spleen samples of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 4 for *Nlrp3*^{-/-}, *n* = 5 for Tau22, *n* = 6 for Tau22/*Nlrp3*^{-/-}. **m-o**, IL-1β, IL-6 and TNF levels in serum samples of 11-month-old wild-type, *Nlrp3*^{-/-}, Tau22 and Tau22/*Nlrp3*^{-/-} mice. *n* = 8 for wild type, *n* = 4 for *Nlrp3*^{-/-} and Tau22, *n* = 6 for Tau22/*Nlrp3*^{-/-}, **P* = 0.0281. All graphs are presented as mean ± s.e.m. and were analysed by one-way ANOVA followed by Tukey's test.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Tau can activate the NLRP3 inflammasome.

a, Immunoblot analysis and quantification of total tau in primary neurons after treatment with conditioned medium from primary wild-type microglia (control), LPS/ATP-activated wild-type, *Asc*^{-/-} or *Nlrp3*^{-/-} knockout microglia (wild type + ATP, *Asc*^{-/-} + ATP or *Nlrp3*^{-/-} + ATP). *n* = 4 for each group. Control versus wild type + ATP: **P* = 0.0252, wild type + ATP versus *Asc*^{-/-} + ATP: **P* = 0.0148, wild type + ATP versus *Nlrp3*^{-/-} + ATP: ***P* = 0.0029. **b**, IL-1 β levels in conditioned medium of primary wild-type microglia primed with LPS and treated with hippocampus homogenate from either 11-month-old wild-type or Tau22 mice. *n* = 5 for primed, *n* = 4 for wild-type and Tau22 homogenate treated microglia. Primed versus Tau22: ***P* = 0.0092, wild type versus Tau22: **P* = 0.0276. **c**, IL-1 β levels in conditioned medium of primary wild-type, *Asc*^{-/-} and *Nlrp3*^{-/-} microglia primed with LPS and treated with different forms of 2 μ M recombinant wild-type tau (tau WT). *n* = 3 for *Asc*^{-/-} and *Nlrp3*^{-/-} microglia treatments and wild-type oligomer treatment, *n* = 8 for all other wild-type treatments. Wild-type primed versus wild-type monomers: ***P* = 0.0011, wild-type primed versus wild-type oligomers: ****P* = 0.0007, wild-type monomers versus *Asc*^{-/-} and *Nlrp3*^{-/-} monomers: ***P* = 0.0011, wild-type monomers versus wild-type fibrils: **P* = 0.0388, wild-type oligomers versus *Asc*^{-/-} and *Nlrp3*^{-/-} oligomers: ****P* = 0.0004, wild-type oligomers versus wild-type fibrils: **P* = 0.0112. **d**, IL-1 β levels in conditioned medium of primary wild-type, *Asc*^{-/-} and *Nlrp3*^{-/-} microglia primed with LPS and treated with different forms of 2 μ M

recombinant tau with a P301S (tau P301S) mutation. *n* = 8 for wild-type microglia treatments, *n* = 3 for *Asc*^{-/-} and *Nlrp3*^{-/-} microglia treatments, ****P* = 0.0002, ***P* = 0.0018. **e**, IL-1 β levels in conditioned medium of primary wild-type microglia primed with LPS and treated with different forms of 2 μ M recombinant tau wild type with and without the NLRP3 inhibitor CRID3. *n* = 4 for all groups, ****P* = 0.0002, *****P* < 0.0001. **f**, IL-1 β levels in conditioned medium of primary wild-type microglia primed with LPS and treated with different forms of 2 μ M recombinant tau P301S with and without CRID3 treatment. *n* = 4 for all groups, ***P* = 0.0037, ****P* = 0.0005, *****P* < 0.0001. **g**, Jess-based analysis of conditioned medium of LPS + tau wild-type-treated wild-type microglia stained for caspase-1. LPS/ATP-treated wild-type microglia served as positive control. **h**, Quantification of data from **g**. *n* = 7 for primed, *n* = 8 for tau monomers and fibrils, *n* = 4 for tau oligomers. **P* = 0.0458, ***P* = 0.0091. **i**, Jess-based analysis of conditioned medium of primary wild-type, *Asc*^{-/-} and *Nlrp3*^{-/-} microglia primed with LPS and treated with the indicated forms of tau P301S. LPS/ATP-treated wild-type microglia served as positive control. Samples were stained for caspase-1. **j**, Quantification of data from **i**. *n* = 7 for primed and fibrils, *n* = 6 for monomers, **P* = 0.0128. For gel source data, see Supplementary Fig. 1. All graphs are presented as mean \pm s.e.m. and were analysed by one-way (**a**, **b**, **h**, **j**) or two-way ANOVA (**c**–**f**) followed by Tukey's test.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	ImageStudio software version 5.2.5 (LI-COR); Compass software version 4.0.0 (ProteinSimple); StepOne software version 2.1 (Applied Biosystems).
Data analysis	ImageStudio software version 5.2.5 (LI-COR); ImageJ/Fiji software version 2.0.0.-rc-67/1.52c; Photoshop CS5 version 12.0.1 (Adobe); Illustrator version 23.0.6 (Adobe); Imaris Version 9 (Oxford Instruments); Compass software version 4.0.0 (ProteinSimple); Bioluminescence express 3D version 3.3; Advanced Analysis 2.0.115 software package; Cytoscape version 3.6.1; Partek Genomics Suite software version 6.6 and 7.18; Interferome data base version 2.01; STRING version 11.0; GraphPad Prism software version 7.0c; StepOne software version 2.1 (Applied Biosystems).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated and/or analyzed during this study are either included in this article (and its supplementary information files) or are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All cell culture experiments requiring statistical analysis were performed at least 3 times as indicated in the figure legends. Power analysis was used to predetermine the sample size in case of in vivo studies. For animal experiments requiring statistical analysis we used at least 5 animals per group with the exception of only 3 animals in the 3 months old WT group used for NanoString analysis. See specific figure legends and extended data figure legends for details.
Data exclusions	All animals were healthy and generated specifically for the experiments described. Animals or samples were excluded from analysis only in the instance of technical failure.
Replication	All experiments were performed with at least three independent biological samples. All attempts at replication were successful.
Randomization	The animals were littermates, and inbred lines were used, where the individual mice were identical, therefore no specific randomization was needed. Mice were grouped according to genotype before they were randomly assigned to the experimental conduct (e.g. Morris Water Maze test).
Blinding	All researchers performing animal experiments and/or data analysis were blinded. However, cell culture treatments and analyses were mostly performed by the same individual, so blinding was not always possible. Wherever possible, a second researcher blinded to the analysis confirmed the result.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	For tissue stainings, anti-biotinylated AT8 (Thermo Fisher Scientific, MN1020B, lot SB2334646, 1:500), anti-Iba1 (Wako, 019-19741, lot PTE0555, 1:400 and Abcam, ab5076, lot GR3245261-1, 1:1500), anti-GFAP (Invitrogen, 13-0300, lot SA247423, 1:100) and anti-pTau-S416 (Abcam, ab119391, lot GR93695-21, 1:1000) were used. For immunoblot analysis, the following antibodies were used: anti-Tau5 (Thermo Fisher Scientific, MA5-12808, lot sc2348223, 1:500), anti-MC1 (gift from P. Davies, New York, USA, 1:1000), anti-PHF-1 (gift from P. Davies, New York, USA, 1:1000), anti-Caspase-1 (Genentech, clone 4B4.2.1, gift from Genentech, San Francisco, CA, 1:1000; Adipogen, clone Bally-1, AG-20B-0048, lot 26101409, 1:1000; Adipogen, clone Casper-1, AG-20B-0042, lot A28881708, 1:50), anti-IL-1beta (Gene Tex, GTX74034, lot 42900, 1:1000), anti-ASC (Adipogen, clone Alz177, AG-25B-0006, lot A40221902, 1:1000), anti-β-actin (Cell Signaling, 4967, lot 11, 1:2000), anti-pCaMKIIα (Cell Signaling, clone D21E4, 12716T, 1:1000), anti-CaMKIIα (Cell Signaling, clone 6G9, 50049S, lot 1, 1:1000), anti-pGSK-3β (BD, clone 13A, 612313, lot 3768, 1:1000), anti-GSK-3β (Cell Signaling, clone 27C10, 9315S, lot 14, 1:1000), anti-p25/p35 (Cell Signaling, clone C64B10, 2680S, lot 5, 1:1000), anti-demPP2A (Merck Millipore, clone 4b7, 05-577, lot 3154938, 1:500), anti-PP2A subunit C (Cell Signaling, clone 52F8, 2259T, lot 2 1:1000) and anti-PME-1 (Merck Millipore, 07-095, lot 2805155, 1:1000). For analysis on the Jess system from ProteinSimple, anti-Caspase-1 (Adipogen, clone Casper-1, AG-20B-0042, lot A28881708, 1:50) was used.
Validation	PHF-1 and MC1 antibodies from Peter Davies are validated for immunoblot analysis several times and widely used in the field. See e.g. Götz et al (2001) Science Aug 24;293(5534):1491-5, Bhaskar et al (2010) Neuron 6;68(1):19-31 and Li and Cho (2019) J Neurochem doi: 10.1111/jnc.14830. Caspase-1 clone 4b4.2.1 from Genetech was validated by using tissue from Caspase-1-

knockout mice. Validation data for commercial antibodies are available on vendor websites. Positive controls were included in every experimental run wherever possible to ensure accurate function of each antibody.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

THY-Tau22 transgenic (Schindowski et al., Am J Pathol 2006), ASC-/- (Millenium Pharmaceuticals, Cambridge, MA, USA) and NLRP3-/- mice (Millenium Pharmaceuticals, Cambridge, MA, USA), all on C57BL/6 genetic background, were used. Male and female animals at the age of 3, 8 and 11 months were used.

Wild animals

Study did not involve wild animals.

Field-collected samples

Study did not involve samples collected from the field.

Ethics oversight

Animal care and handling was performed as approved by the local ethical committees (LANUV NRW 84-02.04.2017.A226).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Post mortem brain material from histologically confirmed AD and FTD cases as well as age-matched controls that had died from non-neurological disease, were derived from the Neurological Tissue Bank of the Biobank of the Hospital Clínic-IDIBAPS. Postmortem times across all cases varied from 6-16,5 hrs. After explantation brain specimen were immediately snap frozen and stored at -80°C until further use. Patients and controls were males and females, 62 ±10 yrs old.

Recruitment

Participants were recruited by the Neurological Tissue Bank of the Biobank of the Hospital Clínic-IDIBAPS.

Ethics oversight

Patients gave their consent to the Biobank of the Hospital-clinic-IDIBAPS. Ethic approval was obtained by the Biobank.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

An interbacterial toxin inhibits target cell growth by synthesizing (p)ppApp

<https://doi.org/10.1038/s41586-019-1735-9>

Received: 13 May 2019

Accepted: 3 October 2019

Published online: 6 November 2019

Shehryar Ahmad^{1,2,9}, Boyuan Wang^{3,9}, Matthew D. Walker², Hiu-Ki R. Tran^{1,2}, Peter J. Stogios^{4,5}, Alexei Savchenko^{4,5,6}, Robert A. Grant³, Andrew G. McArthur^{1,2,7}, Michael T. Laub^{3,8*} & John C. Whitney^{1,2,7*}

Bacteria have evolved sophisticated mechanisms to inhibit the growth of competitors¹. One such mechanism involves type VI secretion systems, which bacteria can use to inject antibacterial toxins directly into neighbouring cells. Many of these toxins target the integrity of the cell envelope, but the full range of growth inhibitory mechanisms remains unknown². Here we identify a type VI secretion effector, Tas1, in the opportunistic pathogen *Pseudomonas aeruginosa*. The crystal structure of Tas1 shows that it is similar to enzymes that synthesize (p)ppGpp, a broadly conserved signalling molecule in bacteria that modulates cell growth rate, particularly in response to nutritional stress³. However, Tas1 does not synthesize (p)ppGpp; instead, it pyrophosphorylates adenosine nucleotides to produce (p)ppApp at rates of nearly 180,000 molecules per minute. Consequently, the delivery of Tas1 into competitor cells drives rapid accumulation of (p)ppApp, depletion of ATP, and widespread dysregulation of essential metabolic pathways, thereby resulting in target cell death. Our findings reveal a previously undescribed mechanism for interbacterial antagonism and demonstrate a physiological role for the metabolite (p)ppApp in bacteria.

Effectors that are exported by the bacterial type VI secretion system (T6SS) are often encoded adjacent to the genes for structural components of the secretion apparatus⁴. In *P. aeruginosa* strain PAO1, the gene for the Tse6 effector is found next to genes encoding the bacteria-targeting haemolysin-coregulated protein secretion island I T6SS (HI-T6SS)⁵. We noted that in the more virulent clinical isolate PA14, a unique domain is encoded by PA14_01140 instead of the well-characterized C-terminal NAD⁺ glycohydrolase toxin domain of Tse6^{6,7} (Fig. 1a). Orthologues of PA14_01140 are found in many PA14-related strains of *P. aeruginosa* as well as several other species of Proteobacteria (Extended Data Fig. 1a, b, Supplementary Dataset 1). An additional open reading frame (PA14_01130) immediately downstream of PA14_01140 may encode a cognate immunity protein, as T6SS effector-immunity genes are typically found adjacent to one another.

We hypothesized that the toxin encoded by PA14_01140 could contribute to the fitness of strain PA14 when co-cultured with PAO1 under contact-promoting conditions that facilitate T6SS attack. Indeed, a PA14 strain lacking PA14_01140 displayed an approximately 40-fold decrease in competitive index against PAO1 (Fig. 1b, Extended Data Fig. 2a). Conversely, a variant of PAO1 lacking *tse6* exhibited a sevenfold decrease in co-culture fitness versus PA14 (Fig. 1c). Although PA14 possesses a homologue of the Tse6-specific immunity determinant *tsi6*, this gene was not protective against Tse6 (Extended Data Fig. 2b). To test the proposed immunity function of PA14_01130, we deleted the

PA14_01140–PA14_01130–*tsi6* gene cluster from a PA14 recipient and found that the recipient was outcompeted by its parental donor strain in a PA14_01140- and HI-T6SS-dependent manner (Fig. 1d, Extended Data Fig. 2c–e). The fitness defect of this recipient was restored by expressing PA14_01130 but not *tsi6* (Fig. 1d). These data demonstrate that PA14_01140 is a T6SS effector and that PA14_01130 is its cognate immunity protein.

We next investigated how PA14_01140 inhibits the growth of bacterial cells. Conventional homology searches were inconclusive, although more sensitive hidden Markov model-based algorithms indicated that the C-terminal domain of PA14_01140 shared weak similarity with proteins harbouring RelA-SpoT homologue (RSH) domains⁸ (Extended Data Fig. 3). These domains are highly conserved across bacteria and usually synthesize the bacterial alarmones guanosine penta- and tetraphosphate, (p)ppGpp, by transferring pyrophosphate from ATP to either GDP or GTP⁹. Intracellular levels of (p)ppGpp regulate growth rate in response to nutritional conditions¹⁰. We found that expression of the C-terminal domain of PA14_01140 (PA14_01140_{tox}) inhibited the growth of *Escherichia coli*, even at levels of approximately three copies per cell, indicating that this domain is sufficient for toxicity (Extended Data Fig. 4a–c).

To determine whether PA14_01140_{tox} is an RSH enzyme, we determined its structure in complex with the PA14_01130 immunity protein to a resolution of 2.2 Å (Fig. 2a, Supplementary Table 1). This structure

¹Michael DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada. ²Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada. ³Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada. ⁵Center for Structural Genomics of Infectious Diseases (CSGID), Toronto, Ontario, Canada. ⁶Department of Microbiology, Immunology and Infectious Diseases, University of Calgary, Calgary, Alberta, Canada. ⁷David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, Ontario, Canada. ⁸Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁹These authors contributed equally: Shehryar Ahmad, Boyuan Wang. *e-mail: laub@mit.edu; jwhitney@mcmaster.ca

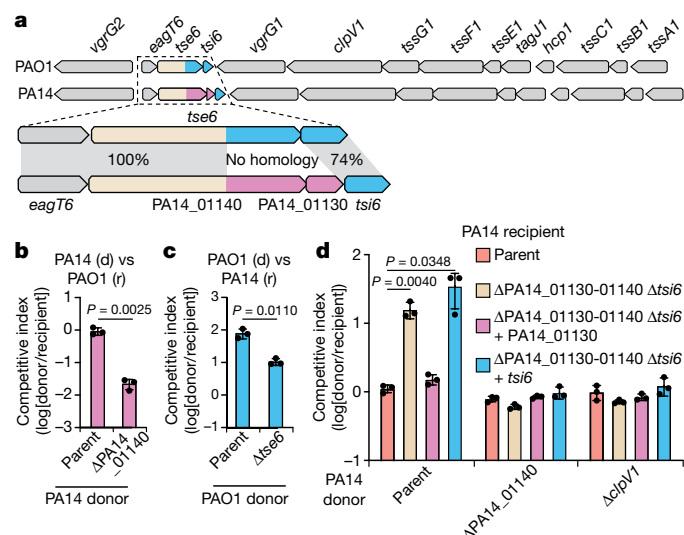


Fig. 1 | A T6SS effector-immunity pair is encoded within the H1-T6SS of *P. aeruginosa* strain PA14. **a**, Genomic context of *tse6-tsi6* and PA14_01140-PA14_01130 within the H1-T6SS gene clusters of *P. aeruginosa* strains PAO1 and PA14, respectively. Known toxin-immunity-encoding regions of *tse6-tsi6* and predicted toxin-immunity-encoding regions of PA14_01140-PA14_01130 are shown in blue and pink, respectively. **b**, **c**, Outcome of growth competition assays between the indicated donor (d) and recipient (r) strains. The parental PA14 genotype is $\Delta rsmA\Delta rsmF$ and the parental PAO1 genotype is $\Delta retS$; both of these mutations stimulate H1-T6SS activity^{20,21}. **d**, Outcome of intraspecific growth competitions between the indicated PA14 donor and recipient strains. The parental PA14 strain genotype is $\Delta rsmA\Delta rsmF$. The competitive index is normalized to starting donor/recipient ratios. **b**-**d**, Mean \pm s.d. for $n = 3$ biological replicates and are representative of two independent experiments; two-tailed, unpaired *t*-tests.

revealed strong similarity to the (p)ppGpp-synthetase domains of RelQ from *Bacillus subtilis* and RelP from *Staphylococcus aureus* (Fig. 2b, Extended Data Fig. 5). Structural overlay of PA14_01140_{tox} with RelQ revealed highly conserved three-dimensional positioning of residues that interact with the pyrophosphate donor ATP (Fig. 2c). Mutation of any of these residues markedly reduced toxicity when mutant PA14_01140_{tox} was expressed in *E. coli* (Extended Data Fig. 4d). In contrast to the ATP binding site, the predicted guanosine nucleotide binding site of PA14_01140_{tox} is substantially distorted relative to the catalytically competent position in Rel enzymes. In our co-crystal structure, two α -helices in PA14_01140_{tox} that were predicted to form this acceptor site are rotated by approximately 30° relative to the equivalent helices in the Rel proteins (Extended Data Fig. 5). This rotation is likely to arise from binding of the immunity protein, PA14_01130, which may neutralize PA14_01140-mediated toxicity by inducing structural rearrangement in the acceptor nucleotide binding site.

A toxin that synthesizes (p)ppApp

To assess the enzymatic activity of PA14_01140_{tox}, we used an assay that couples production of AMP to depletion of NADH, which can be monitored¹¹ at 340 nm. Incubation of purified PA14_01140_{tox} with ATP and GTP led to a dose-dependent decrease in absorbance at 340 nm (A_{340}) over time, indicating that AMP was produced (Fig. 2d). Unexpectedly, production of AMP by PA14_01140_{tox} did not require GTP (Fig. 2e). This finding suggested that PA14_01140_{tox} can transfer a pyrophosphate from ATP to an adenosine nucleotide acceptor. To test this hypothesis, we incubated purified PA14_01140_{tox} with ATP alone, ATP + ADP, or ATP + AMP and used anion-exchange chromatography to identify

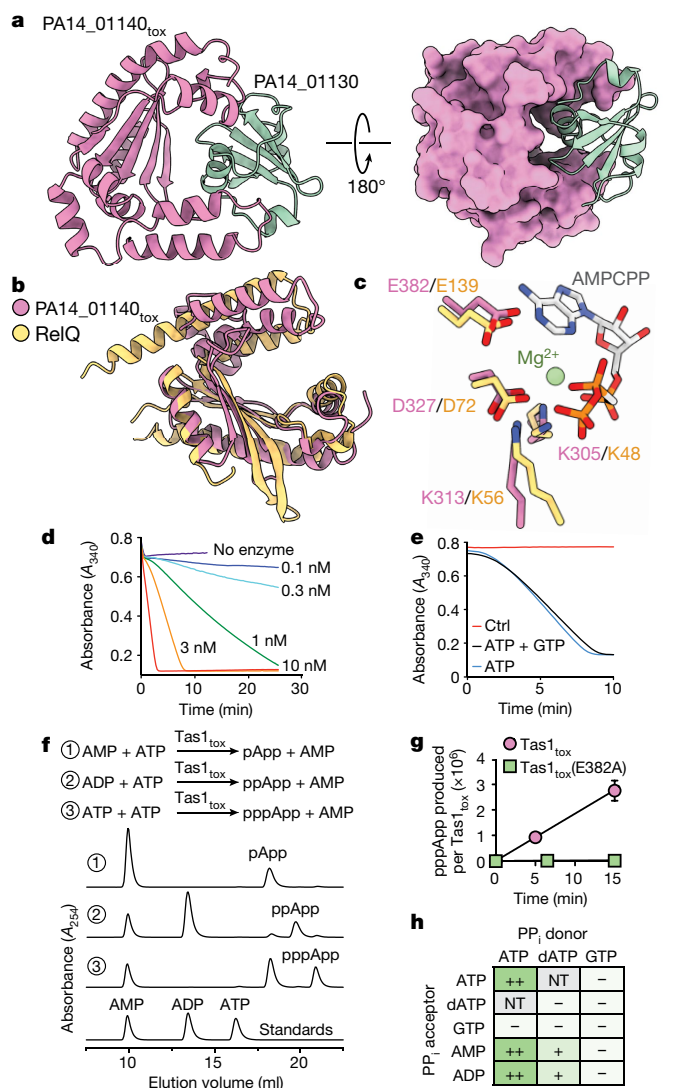


Fig. 2 | Tas1_{tox} adopts an RSH fold found in enzymes that synthesize (p)ppGpp but instead synthesizes (p)ppApp. **a**, Overall structure of PA14_01140_{tox} in complex with PA14_01130. Ribbon (left) and space-filling (right) representations of PA14_01140_{tox} (pink) in complex with PA14_01130 (green). **b**, PA14_01140_{tox} resembles (p)ppGpp synthetase enzymes. Structural overlay of PA14_01140_{tox} and the small alarmone synthetase RelQ from *B. subtilis* (PDB code 5DEC)²². The structures superimpose with a C α root mean square deviation (r.s.m.d.) of 3.4 Å over 145 equivalent positions. **c**, Structural alignment of the pyrophosphate donor ATP binding site of RelQ in complex with a magnesium ion and the non-hydrolysable ATP analogue AMPCPP (PDB code 5F2V) with the equivalent amino acid positions in PA14_01140_{tox}. Amino acid side chains deriving from PA14_01140_{tox} or RelQ and their corresponding labels are shown in pink and yellow, respectively. **d**, PA14_01140_{tox} catalyses the formation of AMP in a dose-dependent manner. Coupled enzyme assay of PA14_01140_{tox}-catalysed AMP production as a function of NADH consumption over time. **e**, PA14_01140_{tox} catalyses the production of AMP from ATP in a GTP-independent manner. The control (ctrl) reaction lacks adenylate kinase, which is required for the initial step of the coupled assay. **f**, PA14_01140_{tox} (Tas1_{tox}) is a (p)ppApp synthetase enzyme. Anion-exchange traces of ATP alone or with excess AMP or ADP after incubation with Tas1_{tox}. A standard trace for ATP, ADP and AMP is shown for comparison. **g**, Rate of production of ppApp by Tas1_{tox} or Tas1_{tox}(E382A). Reactions were performed at 37 °C with 10 mM ATP and 1 nM Tas1_{tox} or 1 μ M Tas1_{tox}(E382A). Mean \pm s.d. from $n = 3$ separate reactions. **h**, Specificity of Tas1_{tox} towards pyrophosphate (PP_i) donors and acceptors. Indicated nucleotides (1 mM each) were incubated with 100 nM Tas1_{tox} at room temperature for 10 min. Reactions that progressed to completion (++), made detectable product (+) or made no detectable product (-) are indicated. NT, not tested (Extended Data Fig. 6). **d**-**f**, Data are representative of two independent experiments.

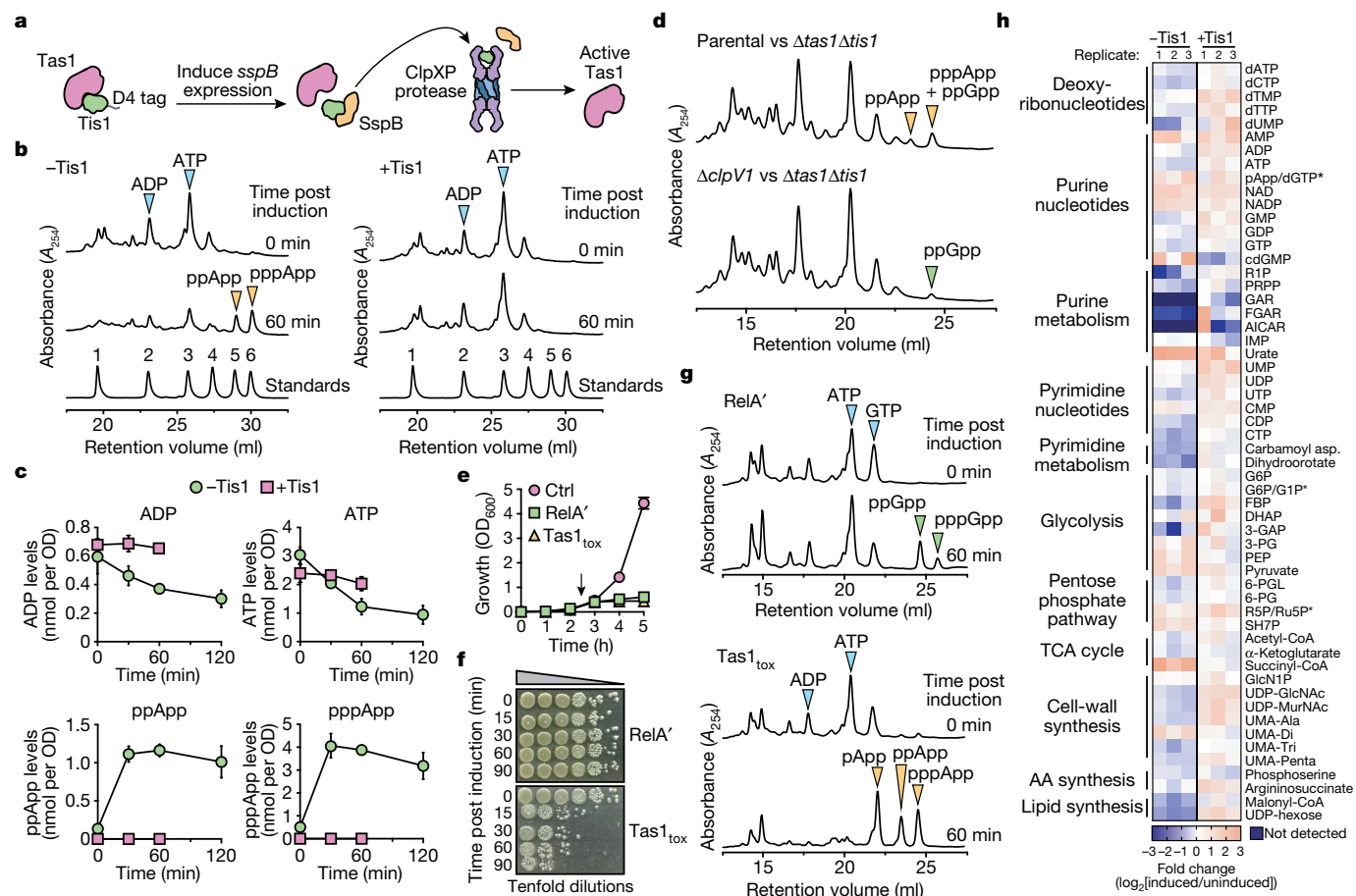


Fig. 3 | Tas1 intoxication depletes cellular ADP and ATP and thereby dysregulates central metabolism. **a**, Schematic of the inducible Tis1 degradation system used to generate active Tas1 in *P. aeruginosa* cells. Induction of *sspB* expression results in degradation of D4-tagged Tis1 by the ClpXP protease²³. **b**, (p)ppApp accumulates in Tis1-depleted *P. aeruginosa* cells. Anion-exchange chromatography separated metabolites extracted from a *P. aeruginosa* PA14 parental strain (right, Δ retS Δ sspB pPSV39-CV::sspB) and a derivative expressing *tis1*-D4 (left, Δ retS Δ sspB PA14_01130-DAS+4 pPSV9-CV::sspB) before or 1 h after induction of *sspB* expression. Blue and orange arrowheads indicate peaks of adenosine 5'-nucleotides and (p)ppApp, respectively. A standard trace of an equimolar mixture of AMP (1), ADP (2), ATP (3), pApp (4), ppApp (5) and pppApp (6) using the same gradient is shown for comparison. **c**, Absolute quantification of ADP, ATP and (p)ppApp in the *P. aeruginosa* strains from **a** as a function of time after induction of Tis1 depletion. OD, optical density. **d**, Anion-exchange chromatography traces of metabolites extracted from growth competition experiments conducted on solid medium for 2.5 h. The parental strain is *P. aeruginosa* Δ rsmA Δ rsmF.

the products. In these reactions, PA14_01140_{tox} produced pppApp, ppApp and pApp, respectively, as verified by mass spectrometry and by ¹H and ³¹P nuclear magnetic resonance (NMR) (Fig. 2f, Extended Data Fig. 6a, b, Supplementary Table 2). Formation of pApp also occurred in the presence of ATP alone, which suggests that the pppApp that is initially produced can subsequently be used to pyrophosphorylate AMP, producing two pApp molecules (Extended Data Fig. 7a, b). Collectively, these results show that PA14_01140_{tox} is a pyrophosphate kinase for adenosine nucleotides. We therefore renamed this effector Tas1 (for type VI secretion effector (p)ppApp synthetase 1) and its cognate immunity protein Tis1 (for type VI secretion immunity to (p)ppApp synthetase 1).

We next examined the catalytic rate of production of pppApp by Tas1. Notably, one molecule of Tas1 could pyrophosphorylate

180,000 molecules of ATP per minute (Fig. 2g). This catalytic rate is two orders of magnitude higher than characterized (p)ppGpp synthetases and is likely to reflect the role of Tas1 as an interbacterial toxin rather than an enzyme involved in controlling growth rate^{12,13}. Turnover was also rapid when ADP or AMP were used as pyrophosphate acceptors (Fig. 2h). In addition to being unable to use GTP as a pyrophosphate donor or acceptor, Tas1 was unable to use dATP as an acceptor, although this deoxynucleotide could serve as a suboptimal pyrophosphate donor (Extended Data Fig. 6c). Substitution of a conserved glutamate residue known to bind the pyrophosphate donor ATP in RSH enzymes with alanine abolished the activity of Tas1 (Fig. 2g, Extended Data Fig. 4d, e).

The remarkable catalytic rate of Tas1 predicts that T6SS-dependent delivery of one toxin molecule into a 1- μ m³ target bacterium would

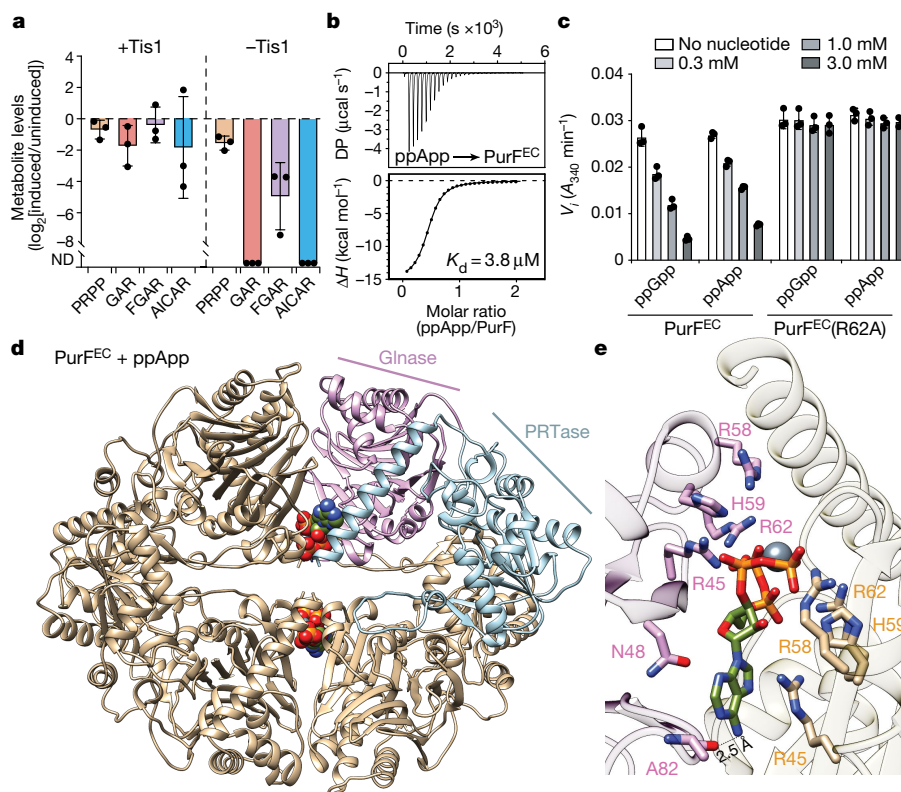


Fig. 4 | (p)ppApp interacts with PurF and inhibits de novo purine biosynthesis. **a**, Relative quantification of metabolites within the de novo purine biosynthesis pathway in *P. aeruginosa* strains containing or lacking Tis1. Metabolite levels for the +Tis1 and -Tis1 strains are shown as \log_2 ratios for samples 1 h after induction relative to before induction of *sspB* expression. ND, not detected. **b**, Isothermal calorimetry trace (top) and fitted isotherm (bottom) for the titration of 100 μ M PurF^{EC} with 1 mM ppApp. A representative trace from two independent experiments is shown. **c**, Changes to the activity of PurF^{EC} in the presence of indicated concentrations of ppGpp or ppApp.

d, Ribbon diagram of the PurF^{EC} tetramer in complex with ppApp. A single PurF subunit is coloured by individual domains (Glnase, glutaminase domain (pink); PRTase, phosphoribosyltransferase domain (blue)); remaining subunits are coloured brown. ppApp and Mg^{2+} are shown in sphere representation. **e**, Close-up view of the ppApp binding site between the glutaminase domains of two adjacent PurF^{EC} monomers. ppApp-interacting residues are shown as pink or brown sticks and hydrogen bonding between PurF^{EC} and the purine ring of ppApp is shown as a black dashed line (Extended Data Fig. 10). **a**, **c**, Mean \pm s.d. for $n = 3$ separate cultures (**a**) or reactions (**c**).

reduce ATP concentration by approximately 0.6 mM per minute. This calculation led us to hypothesize that Tas1 intoxicates cells in part by depleting essential adenosine nucleotides. To test this idea, we first measured nucleotide levels in *E. coli* cells expressing Tas1_{tox}. Within 30 min of expression of Tas1_{tox}, there was a profound reduction in cellular AMP, ADP and ATP levels that coincided with a substantial increase in pApp, ppApp and pppApp (Extended Data Fig. 8a, b).

We also examined nucleotide levels in *P. aeruginosa* cells depleted of the Tis1 immunity protein (Fig. 3a) and observed a similarly large and rapid drop in ADP and ATP levels along with robust formation of pppApp and ppApp (hereafter called (p)ppApp) (Fig. 3b, c). ADP and ATP were not completely depleted, suggesting that intoxicated cells attempt to compensate for the loss of these essential nucleotides by altering their metabolism. AMP levels remained unchanged in Tis1-depleted *P. aeruginosa* cells, which suggests that at physiologically relevant concentrations of Tas1, ADP and ATP are the preferred adenosine nucleotide acceptors. Finally, we detected production of (p)ppApp during interbacterial competition between a PA14 donor and a Δ tas1 Δ tis1 recipient strain in a manner that depended on the presence of a functional T6SS in donor cells (Fig. 3d). Collectively, these results demonstrate that T6SS-delivered Tas1 depletes ADP and ATP in target bacteria by synthesizing (p)ppApp.

To compare the effects of production of (p)ppGpp and (p)ppApp, we assessed the viability of *E. coli* cells expressing either a constitutively active fragment of the (p)ppGpp synthetase RelA (RelA') or Tas1_{tox}.

Even though the expression of both enzymes results in growth arrest, only cells undergoing Tas1-mediated intoxication showed a substantial reduction in viability (Fig. 3e, f, Supplementary Videos 1–4). This difference is likely to arise because, in contrast to (p)ppApp production, production of (p)ppGpp does not significantly reduce ATP levels and results in only a twofold reduction in cellular GTP (Fig. 3g). In line with our findings in *E. coli*, we also observed a decrease in the viability of *P. aeruginosa* cells lacking Tis1 or during interbacterial competition with a Tas1-expressing donor strain (Extended Data Fig. 9a, b). These results indicate that the production of (p)ppGpp is bacteriostatic, whereas the production of (p)ppApp by Tas1 is bactericidal.

(p)ppApp kills target cells in multiple ways

Our findings suggest that (p)ppApp affects target cell physiology by depleting ADP and ATP, which would have pleiotropic consequences for many cellular processes. In particular, ADP is an essential regulator of energy production because of its role in dissipating the proton motive force (pmf) via production of ATP catalysed by ATP synthase. Consequently, reduced levels of ADP following delivery of Tas1 may result in excessive electrostatic potential across the inner membrane. Consistent with this notion, we found that addition of sub-lethal levels of the pmf uncoupling ionophore CCCP reduced the toxicity of Tas1_{tox} (Extended Data Fig. 9c–e). We also tested whether Tas1-intoxicated cells could regenerate ADP by hydrolysing ppApp. In Proteobacteria,

the bifunctional RSH enzyme SpoT can cleave the 3' pyrophosphate of ppGpp to produce GDP¹⁴. We found that the ppGpp-hydrolysing domain of SpoT was substantially less active on ppApp than on ppGpp in vitro (Extended Data Fig. 9f). Furthermore, expression of SpoT in recipient cells lacking Tis1 during interbacterial competition did not result in a change in ppApp levels (Extended Data Fig. 9g). Together, these data suggest that SpoT cannot alleviate Tas1 toxicity by regenerating ADP from ppApp.

Unlike ADP, ATP is required for nearly all anabolic and catabolic pathways in bacteria. To examine the effect of Tas1-dependent depletion of ATP, we performed metabolic profiling of *P. aeruginosa* cells depleted of the Tis1 immunity protein. These Tas1-intoxicated cells displayed a marked decrease in metabolites belonging to many essential pathways, including glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose-phosphate pathway, as well as decreases in intermediates of lipid, amino acid, pyrimidine and purine biosynthesis (Fig. 3h, Supplementary Dataset 2). In addition, the levels of mononucleotide triphosphates and nucleotide-activated precursors involved in cell wall biosynthesis were substantially depleted (Fig. 3h, Supplementary Table 3). Thus, our results suggest that production of (p)ppApp by Tas1 is bactericidal owing to a decrease in ADP and ATP levels, leading to dysregulation of the pmf and depletion of numerous metabolites that are required for cell viability.

We also considered the possibility that (p)ppApp itself contributes to Tas1-mediated toxicity by binding directly to protein targets. As with (p)ppGpp, accumulation of (p)ppApp resulted in a reduction in de novo purine biosynthesis intermediates (Fig. 4a). (p)ppGpp blocks the dedicated step of purine synthesis by competitively inhibiting PurF¹¹. Given its similarity to (p)ppGpp, we hypothesized that (p)ppApp could also inhibit PurF. Indeed, (p)ppApp binds to and inhibits PurF from both *E. coli* (PurF^{EC}) and *P. aeruginosa* (PurF^{PA}) at concentrations of (p)ppApp achieved in Tas1-intoxicated cells (Fig. 4b, c, Extended Data Fig. 10a, b). To determine whether the two nucleotides use similar mechanisms to inhibit PurF, we determined the crystal structure of PurF^{EC} in complex with ppApp to a resolution of 2.5 Å (Supplementary Table 1). Our structure indicated that, despite differences in the purine rings, ppGpp and ppApp bind PurF in a similar manner, and mutation of an arginine residue required for binding of ppGpp to PurF^{EC} also blocked the ability of ppApp to bind and inhibit PurF (Fig. 4c–e, Extended Data Fig. 10c, d). These data indicate that (p)ppApp directly inhibits purine biosynthesis via PurF and that (p)ppApp is likely to target many, if not most, of the more than 50 proteins targeted by (p)ppGpp¹¹, further enhancing the toxicity of (p)ppApp that results from depletion of ADP and ATP.

Conclusions

Our work shows that Tas1 is an interbacterial toxin and also demonstrates delivery of an RSH protein between bacterial cells. In addition, to our knowledge, Tas1 is the first (p)ppApp synthetase enzyme to have a known role in bacterial physiology. All previously characterized RSH enzymes synthesize (p)ppGpp, which regulates cell growth and promotes bacterial survival. Although (p)ppApp is very similar to (p)ppGpp in structure, its physiological role differs because its production irreversibly alters the cellular metabolome, depleting existing pools of ATP and hindering the synthesis of ATP by intoxicated cells. As the *P. aeruginosa* H1-T6SS delivers a diverse payload of effectors into target cells, the (p)ppApp synthetase activity of Tas1 probably augments the activities of co-secreted cell wall and membrane targeting effectors, because pathways involved in cell envelope biosynthesis and repair require ATP^{15–17}.

Although reports from several decades ago linked production of (p)ppApp to sporulation in *B. subtilis*^{18,19}, neither a physiological role for this molecule nor the enzymes that synthesize it had, to our knowledge, been reported. Our discovery of Tas1 indicates that (p)ppApp is a physiologically relevant molecule that can serve as a potent cellular toxin.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1735-9>.

- Granato, E. T., Meiller-Legrand, T. A. & Foster, K. R. The evolution and ecology of bacterial warfare. *Curr. Biol.* **29**, R521–R537 (2019).
- Russell, A. B., Peterson, S. B. & Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. *Nat. Rev. Microbiol.* **12**, 137–148 (2014).
- Haurlyliuk, V., Atkinson, G. C., Murakami, K. S., Tenson, T. & Gerdes, K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nat. Rev. Microbiol.* **13**, 298–309 (2015).
- Wexler, A. G. et al. Human symbionts inject and neutralize antibacterial toxins to persist in the gut. *Proc. Natl Acad. Sci. USA* **113**, 3639–3644 (2016).
- Whitney, J. C. et al. Genetically distinct pathways guide effector export through the type VI secretion system. *Mol. Microbiol.* **92**, 529–542 (2014).
- Whitney, J. C. et al. An interbacterial NAD(P)⁺ glycohydrolase toxin requires elongation factor Tu for delivery to target cells. *Cell* **163**, 607–619 (2015).
- Quentin, D. et al. Mechanism of loading and translocation of type VI secretion system effector Tse6. *Nat. Microbiol.* **3**, 1142–1152 (2018).
- Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
- Atkinson, G. C., Tenson, T. & Haurlyliuk, V. The RelA/SpoT homolog (RSH) superfamily: distribution and functional evolution of ppGpp synthetases and hydrolases across the tree of life. *PLoS One* **6**, e23479 (2011).
- Potrykus, K., Murphy, H., Philippe, N. & Cashel, M. ppGpp is the major source of growth rate control in *E. coli*. *Environ. Microbiol.* **13**, 563–575 (2011).
- Wang, B. et al. Affinity-based capture and identification of protein effectors of the growth regulator ppGpp. *Nat. Chem. Biol.* **15**, 141–150 (2019).
- Gaca, A. O. et al. From (p)ppGpp to (pp)ppGpp: characterization of regulatory effects of pGpp synthesized by the small alarmone synthetase of *Enterococcus faecalis*. *J. Bacteriol.* **197**, 2908–2919 (2015).
- Beljantseva, J. et al. Negative allosteric regulation of *Enterococcus faecalis* small alarmone synthetase RelQ by single-stranded RNA. *Proc. Natl Acad. Sci. USA* **114**, 3726–3731 (2017).
- Sarubbi, E. et al. Characterization of the spoT gene of *Escherichia coli*. *J. Biol. Chem.* **264**, 15074–15082 (1989).
- Buggy, T. D., Braddock, D., Dowson, C. G. & Roper, D. I. Bacterial cell wall assembly: still an attractive antibacterial target. *Trends Biotechnol.* **29**, 167–173 (2011).
- Raetz, C. R. Enzymology, genetics, and regulation of membrane phospholipid synthesis in *Escherichia coli*. *Microbiol. Rev.* **42**, 614–659 (1978).
- LaCourse, K. D. et al. Conditional toxicity and synergy drive diversity among antibacterial effectors. *Nat. Microbiol.* **3**, 440–446 (2018).
- Rhaese, H. J. & Groscurth, R. Control of development: role of regulatory nucleotides synthesized by membranes of *Bacillus subtilis* in initiation of sporulation. *Proc. Natl Acad. Sci. USA* **73**, 331–335 (1976).
- Rhaese, H. J., Hoch, J. A. & Groscurth, R. Studies on the control of development: isolation of *Bacillus subtilis* mutants blocked early in sporulation and defective in synthesis of highly phosphorylated nucleotides. *Proc. Natl Acad. Sci. USA* **74**, 1125–1129 (1977).
- Goodman, A. L. et al. A signaling network reciprocally regulates genes associated with acute infection and chronic persistence in *Pseudomonas aeruginosa*. *Dev. Cell* **7**, 745–754 (2004).
- Marden, J. N. et al. An unusual CsrA family member operates in series with RsmA to amplify posttranscriptional responses in *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **110**, 15055–15060 (2013).
- Steinchen, W. et al. Catalytic mechanism and allosteric regulation of an oligomeric (p)ppGpp synthetase by an alarmone. *Proc. Natl Acad. Sci. USA* **112**, 13348–13353 (2015).
- McGinness, K. E., Baker, T. A. & Sauer, R. T. Engineering controllable protein degradation. *Mol. Cell* **22**, 701–707 (2006).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial strains and growth conditions

P. aeruginosa strains generated in this study were derived from the sequenced strains PAO1 and PA14^{24,25} (Supplementary Table 4). For co-culture experiments, growth curves and secretion assays, *P. aeruginosa* strains were grown at 37 °C in LB medium (10 g/l NaCl, 10 g/l tryptone, and 5 g/l yeast extract). Solid medium contained 1.5% or 3% agar. For analysis of cellular extracts and preparation of samples for metabolomics, *P. aeruginosa* strains were grown at 30 °C overnight and sub-inoculated at 37 °C, in LB medium. Media were supplemented with gentamicin (30 µg/ml) and IPTG (500 µM) as appropriate. *E. coli* strains XL-1 Blue, SM10, BL21 (DE3) CodonPlus, and MG1655 were used for plasmid maintenance, conjugative transfer, gene expression, growth curves and nucleotide extraction experiments, respectively (Supplementary Table 4). *E. coli* strains were grown at 37 °C in LB medium with the exception of the PurF^{PA} expression and nucleotide extraction experiments. For PurF^{PA}, the expression strain was grown in M9 medium (14 g/l Na₂HPO₄·7H₂O, 3 g/l KH₂PO₄, 0.5 g/l NaCl, 1 g/l NH₄Cl, 1 mM MgSO₄, and 30 µM CaCl₂) supplemented with 0.4% glucose and 25 µM Fe(SO₄)-EDTA chelate. For nucleotide extraction experiments, cells were grown in M9 medium supplemented with 0.1% glucose, 0.25% each of L-serine and L-threonine, 0.0375% each of L-asparagine and L-glutamine, 0.015% each of all 16 other natural amino acids, and 1× Kao & Michayluk Vitamin Solution (abbreviated as M9GAV). Where appropriate, media were supplemented with 150 µg/ml carbenicillin, 50 µg/ml kanamycin, 200 µg/ml trimethoprim, 15 µg/ml gentamicin, 500 µM IPTG, 0.1% (w/v) rhamnose or 40 µg/ml X-gal.

DNA manipulation and plasmid construction

All DNA manipulation procedures followed standard molecular biology protocols. Primers were synthesized and purified by Integrated DNA Technologies (IDT). Phusion polymerase, restriction enzymes and T4 DNA ligase were obtained from New England Biolabs (NEB). DNA sequencing was performed by Genewiz Incorporated.

In-frame chromosomal deletion mutants in *P. aeruginosa* were generated using the pEXG2 suicide plasmid as described previously²⁶. In brief, ~500 bp upstream and downstream of target gene were amplified by standard PCR and spliced together by overlap-extension PCR. The resulting DNA fragment was ligated into pEXG2 using standard cloning procedures (Supplementary Table 5). Deletion constructs were introduced into *P. aeruginosa* via conjugal transfer and *sacB*-based allelic exchange was carried out as described previously²⁷. All deletions were confirmed by PCR.

Bioinformatic analysis of *tse6* and *tasI* distribution among *P. aeruginosa* strains

Complete or draft assembled genome sequences for 326 *P. aeruginosa* isolates representing a broad sampling of *P. aeruginosa* diversity were downloaded from the *Pseudomonas* Genome DB²⁸. Open reading frames were predicted for each isolate using Prodigal v2.6.1 and the resulting putative proteomes compared to the Tse6 and TasI sequences using BLASTP v2.8.1, with automated and manual inspection of the results to identify all homologues and sequence variants within each genome^{29,30}. Phylogenetic relationships of the isolates were reconstructed using whole-genome SNP analysis; homologous sites in the genomes containing nucleotide variation among isolates, but not involved in horizontal gene transfer or recombination, were identified using PARSNP v1.2 with PhiPack filtering³¹. The resulting SNP matrix was converted to PHYLIP format and the phylogenetic history of the isolates reconstructed using maximum likelihood as implemented in the RAXML-HPC BlackBox

v8.2.10 hosted on the CIPRES Science Gateway server³². RAXML analysis included automatically generated bootstrapping and estimated proportion of invariables sites (GTRGAMMA+I). The resulting Tse6 and TasI homologues were mapped onto the isolate phylogenetic tree using FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Bacterial toxicity experiments

E. coli XL-1 Blue cells were co-transformed with either pSCrhaB2-CV or pSCrhaB2-CV expressing either wild-type or active site mutants of TasI_{tox} and pPSV39-CV or pPSV39-CV expressing TisI, Tsi6^{PA14} or Tsi6^{PAO1} (Supplementary Table 5). Overnight cultures of these strains were diluted to 10⁶ in tenfold increments and each dilution was spotted onto LB agar plates containing 0.1% (w/v) L-rhamnose, 250 µM IPTG and the appropriate antibiotics. Photographs were taken after overnight growth at 37 °C.

To compare (p)ppApp and (p)ppGpp toxicity, *E. coli* MG1655 with plasmids expressing either TasI_{tox} or RelA' were grown in LB at 30 °C overnight with appropriate antibiotic selection. Stationary-phase overnight cultures were diluted to OD 0.01 in fresh LB medium and grown at 37 °C with shaking. At OD 0.3, 1 ml of culture was retrieved and chilled in ice water and the remaining culture was treated with 500 µM IPTG or 0.1% (w/v) rhamnose. At indicated time points post-induction, 1 ml culture was withdrawn and chilled in ice water for 2 min and then cells were pelleted at 20,000g at 4 °C and re-suspended in ice-cold fresh LB without inducer and kept on ice. These samples were diluted to 10⁶ in tenfold increments and 10 µl from each dilution was spotted onto LB agar plates containing appropriate antibiotics. Nucleotide levels in these strains were examined by anion-exchange chromatography as described below (see 'Metabolite extraction and quantification').

Secretion assay

Stationary-phase overnight cultures of *P. aeruginosa* strains were inoculated in 2 ml LB at a ratio of 1:500. Cultures were grown at 37 °C with shaking to mid-log phase, and cell and supernatant fractions were prepared as described previously³³.

Quantification of TasI_{tox} overexpression in *E. coli*

E. coli MG1655 harbouring an anhydrotetracycline (aTC)-inducible His₆-TasI_{tox}-VSV-G expression plasmid was grown to OD₆₀₀ 0.3 and either untreated or induced with 2 or 3 ng/ml aTC for 15 min. The viability of each culture was assessed by enumerating CFUs. Cells from 100-ml cultures (5 × 10⁹ CFU) were collected for Ni-NTA enrichment of His₆-TasI_{tox}-VSV-G. In brief, cells from each culture were lysed in 1 ml lysis buffer (20 mM HEPES-Na 7.4, 150 mM NaCl) and applied to a column of 0.5 ml Ni-NTA resin. Each column was washed with 3 ml lysis buffer containing 20 mM imidazole followed by 3 ml 20 mM HEPES-Na 7.4, 500 mM NaCl, 20 mM imidazole containing 8 M urea. Bound protein was eluted with a buffer containing 300 mM imidazole and 8 M urea, and each eluate was concentrated to 60 µl. We used 15 µl of concentrated eluate (25% of total protein) for anti-VSV-G immunoblotting for quantification, using 50 and 15 fmol of purified, recombinant His₆-TasI_{tox}-VSV-G as internal standards. Assuming 100% recovery of His₆-TasI_{tox}-VSV-G by Ni-NTA enrichment, cells induced by 2 or 3 ng/ml aTC contain 24 and 44 fmol TasI, which, provided a cell count of 5 × 10⁹, correspond to 3 and 5 copies of His₆-TasI_{tox}-VSV-G per cell, respectively.

Western blot analyses

Western blot analyses of protein samples were performed as described previously using rabbit anti-VSV-G (diluted 1:5,000; Sigma) and rabbit anti-Hcp (diluted 1:5,000) and detected with anti-rabbit horseradish peroxidase-conjugated secondary antibodies (diluted 1:5,000; Sigma)⁶. Western blots were developed using chemiluminescent substrate (Clarity Max, Bio-Rad or SuperSignal West Femto Maximum Sensitivity Substrate, ThermoFisher) and imaged with the ChemiDoc Imaging System (Bio-Rad).

Competition assays

A *lacZ* cassette was inserted into a neutral phage attachment site (*attB*) of recipient *P. aeruginosa* strains to differentiate these strains from unlabelled donors. For interstrain competitions, stationary-phase cultures of *P. aeruginosa* PA14 and PAO1 donor or recipients were mixed in a 4:3 (v/v) ratio, in favour of the donor. For intraspecific competitions, stationary-phase cultures of *P. aeruginosa* PA14 donor and recipient strains were mixed in a 1:1 (v/v) ratio.

Starting ratios of donor and recipient were enumerated by plating on LB agar containing 40 µg/ml X-gal. Ten microlitres of each competition mixture was then spotted in triplicate on a 0.45-µm nitrocellulose membrane overlaid on a 3% LB agar plate and incubated face up at 37 °C for 20–24 h. Competitions were then harvested by resuspending cells in LB and enumerating CFUs by plating on LB agar containing 40 µg/ml X-gal. The final ratio of donor/recipient CFUs was normalized to the starting ratio of donor and recipient strains.

To monitor (p)ppApp production in recipients by anion exchange, 600 µl of donor and recipient mixtures were plated on a 25-mm, 0.45-µm PVDF membranes using vacuum filtration. The membrane was overlaid onto 3% LB agar and incubated face up for 2–7 h. Following incubation, each membrane was immersed in 2 ml ice-cold lysis solvent, a methanol–acetonitrile–water mixture in a volume ratio of 40:40:20. After brief sonication to detach cells from the PVDF membrane, the membrane was removed, and the entire suspension was diluted into 6 ml of 20 mM Tris-HCl buffer (pH 8.0). The diluted mixture was spun at 10,000g to pellet insoluble debris. After passage of a 0.22-µm syringe filter, 4 ml of the supernatant was analysed using anion-exchange chromatography as described below (see Metabolite extraction and quantification).

Tis1 depletion system

A C-terminal *ssrA*-like DAS+4 degradation tag (peptide sequence of tag: AANDENYSENADAS) was fused to the 3' end of the native *tis1* locus in a *P. aeruginosa* strain bearing deletions in *retS* and *sspB*²³. An IPTG-inducible plasmid containing *sspB* was used to stimulate controlled degradation of Tis1–DAS+4 (Tis1–D4). The SspB protein recognizes DAS+4 tagged proteins and delivers them to the ClpXP protease for degradation.

Protein expression and purification

Tas1_{tox}–Tis1 complex. Tas1_{tox} or Tas1_{tox} (E382A) was coexpressed with Tis1 from pETDuet-1 in *E. coli* BL21 (DE3) CodonPlus cells. Forty-millilitre overnight cultures of expression strains were diluted into 2 l of LB broth and grown to mid-log phase (OD₆₀₀ 0.6) in a shaking incubator at 37 °C. Protein expression was induced by the addition of 1 mM IPTG and cells were further incubated for 3.5 h at 37 °C. Cells were harvested by centrifugation at 9,800g for 10 min and resuspended in 25 ml lysis buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 10 mM imidazole) before rupture by sonication (6 × 30-s pulses, amplitude 30%). Cell lysates were cleared by centrifugation at 39,000g for 60 min and the soluble fraction was loaded onto a gravity flow Ni-NTA column that had been equilibrated in lysis buffer.

To obtain Tas1_{tox}–Tis1 complex for crystallization, Ni-NTA bound complex was washed twice with 25 ml lysis buffer followed by elution in 10 ml lysis buffer supplemented with 400 mM imidazole. The Ni-NTA purified complex was further purified by gel filtration using a HiLoad 16/600 Superdex 200 column equilibrated in 20 mM Tris-HCl pH 8.0 150 mM NaCl. Fractions with the highest purity were used for subsequent crystallization screening.

To obtain Tas1_{tox} and Tas1_{tox} (E382A) for enzyme assays, Tis1 was removed from Ni-NTA-immobilized Tas1_{tox} or Tas1_{tox} (E382A) by washing the column twice with 25 ml lysis buffer supplemented with 8 M urea. On-column refolding was achieved by washing twice with 25 ml lysis buffer followed by elution of the renatured proteins using lysis

buffer supplemented with 400 mM imidazole. Refolded Tas1_{tox} and Tas1_{tox} (E382A) were further purified by gel-filtration as described above, except that the running buffer was comprised of 20 mM HEPES pH 7.4, 150 mM NaCl. Purified proteins were then flash frozen until needed.

SpoT₁₋₃₈₇. The SpoT₁₋₃₈₇ fragment from *P. aeruginosa* was expressed from pETDuet-1 in *E. coli* BL21 (DE3) CodonPlus cells. The same expression protocol was followed as described for the Tas1_{tox}–Tis1 complex. To obtain SpoT₁₋₃₈₇ for enzyme assays, cleared cell lysates containing SpoT₁₋₃₈₇ were loaded onto a gravity flow Ni-NTA column that had been equilibrated in lysis buffer. The Ni-NTA bound SpoT₁₋₃₈₇ was washed twice with 25 ml lysis buffer followed by elution in 10 ml lysis buffer supplemented with 400 mM imidazole. The Ni-NTA purified complex was further purified by gel-filtration as described above except that the running buffer was comprised of 20 mM HEPES pH 7.4, 150 mM NaCl. Purified proteins were then flash frozen until needed.

PurF^{EC} and PurF^{PA}. PurF^{EC} used for crystallization was expressed without an affinity tag (Supplementary Table 5). PurF^{EC} and PurF^{PA} used in biochemical experiments were expressed as fusion proteins with a C-terminal self-cleaving Cfa-His₆ tag³⁴. Cultures of expression strains were grown to mid-log phase, cooled to 22 °C and induced with 200 µM IPTG for 20 h. For PurF^{PA}, M9 minimal medium was used for expression.

For untagged PurF^{EC}, cell pellets (~10 g wet weight) were resuspended in 40 ml lysis buffer containing 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM MgCl₂, 5 mM DTT, 20 µg/ml lysozyme and 1 mM PMSF. Cells were lysed by sonication and centrifuged at 15,000g for 10 min. Cleared lysates were treated with protamine sulfate (8 mg per gram of cell pellet) and vortexed. Precipitate was pelleted at 30,000g for 1 h and cleared lysate was fractionated using a DEAE sepharose column equilibrated in buffer A (50 mM Tris-HCl, pH 8.0, 10 mM MgCl₂ and 5 mM DTT). The column was washed with 50 ml 5% buffer B (buffer A + 1 M NaCl) and bound protein was eluted using a linear gradient with buffer B concentration increasing from 5% to 55% over 200 ml. Peak fractions were combined and saturated ammonium sulfate was added to samples at 4 °C. Precipitated protein collected between 40% and 47.5% saturation was redissolved in gel-filtration buffer (20 mM HEPES-Na, pH 7.4, 150 mM NaCl, 2 mM MgCl₂ and 1 mM TCEP) and run over a Superdex-200 increase (10/300) column.

PurF–Cfa-His₆ was purified using a Ni-NTA affinity column, as previously described. To cleave the Cfa-His₆ tag, eluate was treated with 100 mM sodium 2-mercaptoethanesulfonate (MESNa), 100 mM L-cysteine, and 20 mM TCEP, pH 7.0 at room temperature overnight. The cleavage mixture was dialysed against gel filtration buffer and then subjected to a reverse Ni-NTA process. Collected protein was run over a Superdex-200 column equilibrated with gel-filtration buffer.

Crystallization and structure determination

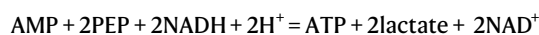
Tas1_{tox}–Tis1 complex. Selenomethionine-incorporated Tas1_{tox}–Tis1 complex was concentrated to 28 mg/ml by spin filtration (10 kDa MWCO, MilliporeSigma) and screened for crystallization conditions at 23 °C using commercially available sparse matrix screens (MCSG1-4, Anatrace) and the hanging drop vapour diffusion technique. Diffraction-quality crystals appeared after approximately one week in a condition containing sodium acetate pH 4.5, 0.8 M sodium phosphate monobasic, 1.2 M potassium phosphate dibasic. Crystals were cryoprotected in crystallization buffer containing 20% (v/v) ethylene glycol and flash frozen in liquid nitrogen before data collection. Synchrotron diffraction data were collected at 100K on a Pilatus 3 × 6M detector using beamline 19-ID of the Structural Biology Center at the Advanced Photon Source. Diffraction data were processed using HKL3000³⁵ and the structure was solved by single wavelength anomalous dispersion (SAD) phasing using Phenix.autosol³⁶. Initial model building was performed using Phenix.autobuild followed by manual adjustment

using Coot³⁷. Refinement was carried out using Phenix.refine with TLS parameterization.

PurF^{EC}-ppApp complex. Crystals were grown by hanging-drop vapour diffusion with drops containing 2 µl protein (25 mg/ml PurF^{EC} in 20 mM HEPES-Na, pH 7.4, 150 mM NaCl, 1 mM TCEP, 5 mM ppApp, and 10 mM MgCl₂) mixed with 2 µl well solution (0.1 M HEPES-Na, pH 7.4, 24% PEG 3350 and 4% iPrOH) at 18 °C. After 1 week, crystals were flash frozen in liquid nitrogen without added cryoprotectant. Diffraction data were collected at the APS, with the NE-CAT beamline 25-IDC on a Pilatus 6M detector. Diffraction data were indexed, integrated and scaled using XDS/XSCALE³⁸ and refined with Phenix³⁶. The structure was solved by molecular replacement using Phaser³⁹ with chain A of PDB entry 6CZF as the search model. The asymmetric unit of the C222₁ cell contains two PurF chains forming a symmetric dimer. The D2 symmetry of the PurF tetramer is generated by the crystallographic centring operation. As in the 6CZF crystal, each PurF tetramer has four ligand binding sites but can bind only two ligands because pairs of binding sites overlap each other across a twofold symmetry axis of the tetramer. Consequently, each PurF chain in the PurF-ppApp crystal is modelled with a single ppApp (and its associated Mg²⁺ ion) at 0.5 occupancy.

Biochemical analysis of Tas1_{tox}

Analysis by coupled enzyme assay. Each reaction (100 µl) contained 50 mM HEPES 7.4, 150 mM NaCl, 20 mM KCl, 10 mM MgCl₂, 1 mM TCEP-Na, 5 mM ATP, 1 mM GTP (if indicated) and Tas1_{tox} or Tas1_{tox}(E382A) at the indicated concentrations. To couple production of AMP in the pyrophosphokinase reaction to the consumption of NADH, the reaction also contained 3.75 M phosphoenolpyruvate (PEP), 0.5 mM NADH, 10 U/ml myokinase (adenylate kinase, ADK), 20 U/ml pyruvate kinase (PK) and 20 U/ml lactate dehydrogenase (LDH).



Reactions were assembled in 96-well plates, with Tas1 added at $t = 0$. The reactions were monitored at 25 °C in a Spectramax M5 plate reader (Molecular Devices) and absorbance at 340 nm (A_{340}) was measured every 15 s.

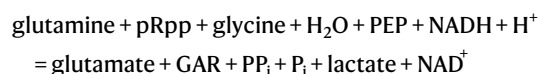
Analysis by anion-exchange chromatography. Each reaction (100 µl total volume) contained 20 mM HEPES-Na 7.4, 300 mM NaCl, 10 mM MgCl₂, and substrates at indicated concentrations. Tas1_{tox} was diluted to 10× working concentration in the above buffer conditions and added last. Reactions were incubated at 37 °C (Tas1_{tox} turnover experiment in Fig. 2g) or 25 °C (all other reactions). At the indicated time points, each 50-µl reaction was diluted in 1 ml ice-cold water and then applied to a MonoQ 5/50 column (GE Healthcare). Bound nucleotides were eluted at 4 °C using a linear gradient of buffer A (5 mM Tris-HCl pH 8.0) and buffer B (5 mM Tris-HCl pH 8.0, 1M NaCl), with the percentage of buffer B increasing from 0 to 40% over 20 ml.

Tas1 turnover measurement. Each reaction (200 µl total volume) contained 20 mM HEPES-Na pH 7.4, 150 mM NaCl, 15 mM MgCl₂, 10 mM ATP, 25 mM PEP-K, 10 U/ml each ADK and PK, and 1 nM Tas1_{tox} or 1 µM Tas1_{tox}(E382A). Reactions were incubated at 37 °C and 20 µl was diluted in 1 ml ice-cold water at the indicated time points and analysed by anion-exchange chromatography as described above. The 3'-pyrophosphorylated product, pppApp, was quantified based on the integration of the A₂₅₄ trace.

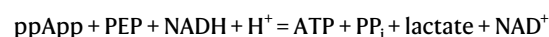
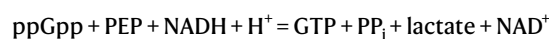
Additional biochemical analyses by enzyme-coupling read-out

Reactions were assembled in 96-well plates from 10× stocks of individual components, with enzymes added at $t = 0$. Then reactions were monitored at 30 °C in a Spectramax M5 plate reader (Molecular Devices) for the A₃₄₀ every 30 s.

PurF glutamine amidophosphoribosyltransferase assay. Each reaction (100 µl) contained 50 nM PurF^{EC} or 100 nM PurF^{PA} in 50 mM HEPES-Na pH 7.4, 150 mM NaCl, 10 mM MgCl₂, 1 mM TCEP, 5 mM glutamine, 1 mM pRpp-Mg and indicated concentrations of ppGpp-Mg, ppApp-Mg or pppApp-Mg. To couple production of 5'-phosphoribosylamine (PRA) by PurF to the consumption of NADH, the reaction also contained 5 mM ATP, 5 mM glycine and 1 µM *E. coli* PurD (these components ligate glycine to 5'-PRA to form glycinamide 5'-ribonucleotide (GAR) and generate ADP), as well as 3.75 M PEP, 0.5 mM NADH, 20 U/ml PK and 20 U/ml LDH.



***P. aeruginosa* SpoT₁₋₃₈₇ hydrolase assay.** Each reaction (100 µl) contained 40 mM HEPES 7.4, 150 mM KCl, 5 mM MnCl₂, 2 mM MgCl₂, 1 mM TCEP-Na, 1 mM ATP, 10 µM SpoT₁₋₃₈₇, and 1 mM ppGpp or ppApp. To couple production of ADP or GDP by SpoT₁₋₃₈₇ to the consumption of NADH, the reaction also contains 3.75 M PEP, 0.5 mM NADH, 1 µM *E. coli* nucleoside diphosphate kinase, 20 U/ml PK and 20 U/ml LDH.



Scale-up preparation of (p)ppApp

pApp. Because Tas1 can slowly convert pppApp and AMP to pApp, quantitative conversion of ATP to pApp is achieved after prolonged incubation with Tas1. Excess AMP was included to ensure the complete consumption of pppApp. We incubated 175 µmol AMP and 75 µmol ATP with 50 nmol Tas1 in 50 ml reaction containing 20 mM HEPES 7.4, 150 mM NaCl, 5 mM MgCl₂ and 1 mM TCEP-Na. The reaction reached completion after 15 min at room temperature and then diluted to 200 ml with iced water. pApp was purified using anion-exchange chromatography.

ppApp. When ADP is present in excess to ATP, Tas1 activity preferentially produces ppApp (Fig. 2f). Thus, we first incubated 50 µmol ADP with 50 pmol Tas1_{tox} in 5 ml 20 mM HEPES 7.4, 150 mM NaCl, 20 mM MgCl₂ and 1 mM TCEP-Na at 37 °C. Then, with vigorous stirring, we added 45 µmol ATP in nine portions over 10 min. After another 5 min of incubation at 37 °C, the reaction was complete and Tas1_{tox} was inactivated with 2 ml chloroform. The aqueous phase was isolated and diluted to 25 ml with water, and ppApp was purified using anion-exchange chromatography.

pppApp. After synthesizing pppApp, Tas1_{tox} further converts pppApp and AMP into pApp (Extended Data Fig. 7). To maximize the yield of pppApp, we included ADK, PK and PEP to regenerate ATP from AMP. The synthesis was thus carried out with 50 µmol ATP, 125 µmol PEP, 25 pmol Tas1_{tox} and 200 U/ml each PK and ADK in 5 ml in the presence of 20 mM HEPES 7.4, 150 mM NaCl, 15 mM MgCl₂ and 1 mM TCEP-Na. After incubation at 37 °C for 30 min, 250 nmol Tas1_{tox} was added and the mixture was incubated for another 30 min. Tas1_{tox} was then inactivated with 2 ml chloroform. The aqueous phase was isolated and diluted to 25 ml with water, and pppApp was purified using anion-exchange chromatography.

Preparative anion-exchange chromatography. To purify (p)ppApp, a MonoQ 10/100 column (GE Healthcare) was operated at 5 ml/min at room temperature. (p)ppApp synthesis reactions were diluted with water and applied to the column. Bound nucleotides were eluted using a linear gradient of buffer A (5 mM Tris-HCl pH 8.0) and buffer B (5 mM Tris-HCl pH 8.0, 1M NaCl), with the percentage of buffer B increasing

Article

from 15 to 40% within 5 column volumes (~40 ml). Preparations of ppApp were purified in two runs, while preparations of ppApp were purified in four runs. Fractions containing the purified product were combined, and LiCl was added to the combined fractions to 1 M final concentration. Then, 4× volumes of ethanol were added to precipitate the nucleotide. After incubation in an ice-water bath for 30 min, the nucleotide was collected by centrifugation at 8,000g for 10 min and the mother liquor decanted. The product was washed with 10 ml 95% ethanol, then dissolved in water and dried on a lyophilizer. The powder was reconstituted in water and concentration determined by absorbance at 260 nm ($\epsilon = 15,400 \text{ M}^{-1} \text{ cm}^{-1}$).

Metabolite extraction and quantification

Culture and induction conditions. Prior to each experiment, strains were grown overnight at 30 °C to stationary phase in the same medium. The starter culture was diluted to OD_{600} 0.005 in fresh medium and grown at 37 °C. Inducer was added after OD_{600} reached 0.10 for *P. aeruginosa* or 0.25 for *E. coli*. Untreated control samples were harvested 1 min before induction.

For expression of Tas1_{tox} and Tas1_{tox} (E382A) in *E. coli* MG1655, cells were grown in LB or M9GAV containing 250 µg/ml trimethoprim (TMP) and induced using 0.1% rhamnose.

For expression of RelA' in *E. coli* MG1655, cells were grown in LB containing 100 µg/ml carbenicillin and induced using 500 µM IPTG.

For depletion of Tis1 in *P. aeruginosa* PA14, the Tis1 inducible-degradation strain *retSΔsspB* PA14_01130-DAS+4 pPSV9-CV::sspB and its parental strain *ΔretSΔsspB* pPSV9-CV::sspB were grown in LB containing 50 µg/ml gentamycin and Tis1 depletion was induced using 500 µM IPTG.

Metabolite extraction from *E. coli*. *E. coli* cells (2.5–3.5 OD) were collected on a 0.22-µm hydrophilic PVDF membrane by vacuum filtration and washed briefly with 160 mM NaCl. At the same time, the culture was sampled for OD_{600} measurements. Cells on the membrane were subsequently immersed in ice-cold lysis solvent, a methanol–acetonitrile–water mixture in a volume ratio of 40:40:20. Lysates were briefly sonicated and, after removal of PVDF membranes, diluted by the lysis solvent for a uniform cell density, typically 1.0 OD_{600} cells per ml solvent.

Metabolite extraction from *P. aeruginosa*. *P. aeruginosa* cells (1.25–1.75 OD) were collected on a 0.45-µm hydrophilic PVDF membrane by vacuum filtration and washed briefly with 160 mM NaCl. At the same time, the culture was sampled for OD_{600} measurements. Cells on the membrane were subsequently immersed in ice-cold lysis solvent, a methanol–acetonitrile–water mixture in a volume ratio of 40:40:20 containing 0.02% (v/v) metabolomics amino acid mix standard solution (Cambridge Isotope Laboratories, MSK-A2-1.2) as the internal standard (ISTD). After brief sonication to detach cells from the PDVF membrane, the membrane was removed, and the suspension was diluted using the lysis solvent to 0.625 OD_{600} cells per ml solvent.

Nucleotide quantification using anion-exchange chromatography. (p)ppApp was quantified using anion-exchange chromatography. In brief, a cell suspension in lysis solvent equivalent to 1.0 OD_{600} cells was diluted with aqueous solution of 10 mM Tris-HCl pH 8.0 until the content of organic solvent was less than 20%. Insoluble material was pelleted at 10,000g, and the supernatant was applied to a Mono Q 5/50 column (GE Healthcare) after passing through a 0.22-µm syringe filter. Bound metabolites were eluted at 4 °C using a linear gradient of buffer A (5 mM Tris-HCl pH 8.0) and buffer B (5 mM Tris-HCl pH 8.0, 1M NaCl), with the percentage of buffer B increasing from 0 to 35% over 17.5 ml. External standards containing equimolar AMP, ADP, ATP, pApp, ppApp and pppApp were analysed under the same conditions to locate their peaks. Nucleotides were quantified according to their peak areas on the 254-nm chromatogram.

MS profiling of *P. aeruginosa* metabolites. Cell suspension in lysis solvent (0.625 OD/ml) was extracted with 1.5× volumes of water and cell debris removed by centrifugation. Cleared extract (330 µl) was mixed with 770 µl 50% methanol in acetonitrile (v/v), and the mixture was frozen at –40 °C for 1 h. Any insoluble material was spun down at 4 °C, 20,000g for 10 min. One millilitre of supernatant (0.075 OD) was transferred to a fresh tube and solvent evaporated using a speedvac followed by a lyophilizer. The residual was reconstituted with 37.5 µl water, and 4 µl was injected into a ZIC-pHILIC 150 × 2.1 mm (5 µm particle size) column (EMD Millipore). Analysis was conducted on a QExactive benchtop orbitrap mass spectrometer equipped with an Ion Max source and a HESI II probe, which was coupled to a Dionex UltiMate 3000 UPLC system (Thermo Fisher Scientific). External mass calibration was performed using the standard calibration mixture every seven days. Chromatographic separation was achieved using the following conditions: buffer A was 20 mM ammonium carbonate, 0.1% ammonium hydroxide; buffer B was acetonitrile. The column oven and autosampler tray were held at 25 °C and 4 °C, respectively. The chromatographic gradient was run at a flow rate of 0.150 ml/min as follows: 0–20 min: linear gradient from 80% to 20% B; 20–20.5 min: linear gradient from 20% to 80% B; 20.5–28 min: hold at 80% B. The mass spectrometer was operated in full-scan, polarity switching mode with the spray voltage set to 3.0 kV, the heated capillary held at 275 °C, and the HESI probe held at 350 °C. The sheath gas flow was set to 40 units, the auxiliary gas flow was set to 15 units, and the sweep gas flow was set to 1 unit. MS data acquisition was performed in a range of 70–1,000 *m/z*, with the resolution set at 70,000, the AGC target at 10×10^6 , and the maximum injection time at 20 ms. Relative quantification of polar metabolites was performed with XCalibur QuanBrowser 2.2 (Thermo Fisher Scientific) using a 5 ppm mass tolerance and referencing an in-house library of chemical standards.

For relative quantifications, a peak area of 1.0×10^4 was arbitrarily assigned to undetected metabolites. Then, peak area of each metabolite was normalized to the ISTD amino acid with the closest retention time and ionized by the same charge. Fold change of metabolite levels between conditions were then calculated based on normalized peak areas.

For absolute quantifications, standard samples containing AMP, ADP, ATP, GMP, GDP, GTP, pApp, ppApp, pppApp, IMP, UTP, dATP, dGTP, dCTP, dTTP and UDP-GlcNAc at a series of known concentrations were prepared in water containing 0.064% (v/v) ISTD and 4 µl was analysed under the same conditions. Note that this ISTD concentration was identical to that in metabolome samples. Peak areas of all 16 nucleotides were therefore normalized to that of $^{13}\text{C}_5$ - ^{15}N -glutamate, and standard curves were generated. Absolute levels of the above nucleotides in unknown samples were then derived through interpolation.

Isothermal titration calorimetry

All isothermal titration calorimetry (ITC) experiments were performed in a VP-ITC (Malvern) instrument thermo-equilibrated at 25 °C with water in the reference cell. Ligand solution was 1 mM (p)ppApp and 1 mM MgCl₂ in a buffer containing 20 mM HEPES-Na pH 7.4 and 150 mM NaCl, 2 mM MgCl₂ and 1 mM TCEP. The sample cell contained 100 µM PurF^{EC} in the same buffer. ppApp-Mg was injected in 27 injections at 10 nmol/injection. Blank titrations were performed with protein-free gel filtration buffer in the sample cell. The blank-subtracted data were analysed using the Origin software package (version 5.0, MicroCal) and fit using a single-site binding model.

Microscopy

Phase contrast and propidium iodide-fluorescence images were taken on a Zeiss Observer Z1 microscope using a 100×/1.4 oil immersion objective and an LED-based Colibri illumination system using software Metamorph (Universal Imaging). Cells were first washed with inducer-free medium

and concentrated to OD₆₀₀ 0.5. A 1- μ l sample was spotted onto 1.5% agarose LB pads containing 2.5 μ g/ml propidium iodide and incubated at 30 °C. Time-lapse images were taken every 10 min over a 6-h period.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data supporting the findings of this study are available within the manuscript and associated Supplementary Information. X-ray crystallographic coordinates and structure factor files are available from the PDB with the following accession numbers: Tas1_{tox}-Tis1 (6OX6) and PurF^{EC}-ppApp (6OTT). Maximum likelihood estimates of *P. aeruginosa* strain relationships used for tree construction are provided in Newick format in Supplementary Dataset 1. Relative concentrations of metabolites from metabolomics are reported in Supplementary Dataset 2. Source gel images are available in Supplementary Fig. 1.

24. Stover, C. K. et al. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).
25. Lee, D. G. et al. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol.* **7**, R90 (2006).
26. Rietsch, A., Vallet-Gely, I., Dove, S. L. & Mekalanos, J. J. ExsE, a secreted regulator of type III secretion genes in *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **102**, 8006–8011 (2005).
27. Hmelo, L. R. et al. Precision-engineering the *Pseudomonas aeruginosa* genome with two-step allelic exchange. *Nat. Protocols* **10**, 1820–1841 (2015).
28. Winsor, G. L. et al. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Res.* **44**, D646–D653 (2016).
29. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
30. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
31. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524 (2014).
32. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
33. Hood, R. D. et al. A type VI secretion system of *Pseudomonas aeruginosa* targets a toxin to bacteria. *Cell Host Microbe* **7**, 25–37 (2010).

34. Stevens, A. J. et al. Design of a split intein with exceptional protein splicing activity. *J. Am. Chem. Soc.* **138**, 2162–2165 (2016).
35. Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr. D* **62**, 859–866 (2006).
36. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
37. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
38. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
39. Bunkóczi, G. et al. Phaser.MRage: automated molecular replacement. *Acta Crystallogr. D* **69**, 2276–2286 (2013).
40. Manav, M. C. et al. Structural basis for (p)ppGpp synthesis by the *Staphylococcus aureus* small alarmone synthetase RelP. *J. Biol. Chem.* **293**, 3254–3264 (2018).

Acknowledgements We thank A. Raphenya and B. Alcock for assistance with sequence data curation and analyses, C. Chang for assistance with X-ray data collection and processing, and the Whitehead Institute Metabolite Profiling Core Facility for measuring metabolite levels. S.A. and B.W. were supported by an Ontario Graduate Scholarship and a fellowship from the Jane Coffin Childs Memorial Fund, respectively. A.G.M. holds a Cisco Research Chair in Bioinformatics and M.T.L. is an Investigator of the Howard Hughes Medical Institute. Results shown in this report are derived from work performed by the Structural Biology Center (SBC) and the Northeastern Collaborative Access Team (NECAT) at the Advanced Photon Source, Argonne National Laboratory. SBC is funded by NIAID (HHSN272201200026C) and HHS (HHSN272201700060C) and NECAT is funded by NIH grants P30 GM124165 and S10OD021527. SBC-CAT and NECAT are operated by UChicago Argonne, LLC, for the US DOE under contract number DE-AC02-06CH11357. This work was supported by grants from the Canadian Foundation for Innovation (34531 to A.G.M.), NIH (R01-GM082899 to M.T.L.) and CIHR (PJT-156129 to J.C.W.), and by seed funding from the David Braley Centre for Antibiotic Discovery (to J.C.W.).

Author contributions Experiments were conceived and designed by S.A., B.W., M.T.L. and J.C.W. Cloning, bacterial competition assays, protein purification, biochemical experiments and protein crystallization were carried out by S.A. and B.W. X-ray data collection and analyses were performed by P.J.S. and R.A.G. Bioinformatics analyses for Extended Data Fig. 1a were performed by H.-K.R.T. and A.G.M. Assistance with cloning, purification and crystallization of Tas1-Tis1 complex was provided by M.D.W. Figure design, manuscript writing and editing were done by S.A., B.W., M.T.L. and J.C.W. The project was supervised by M.T.L. and J.C.W. Funding was provided by A.S., R.A.G., A.G.M., M.T.L. and J.C.W.

Competing interests The authors declare no competing interests.

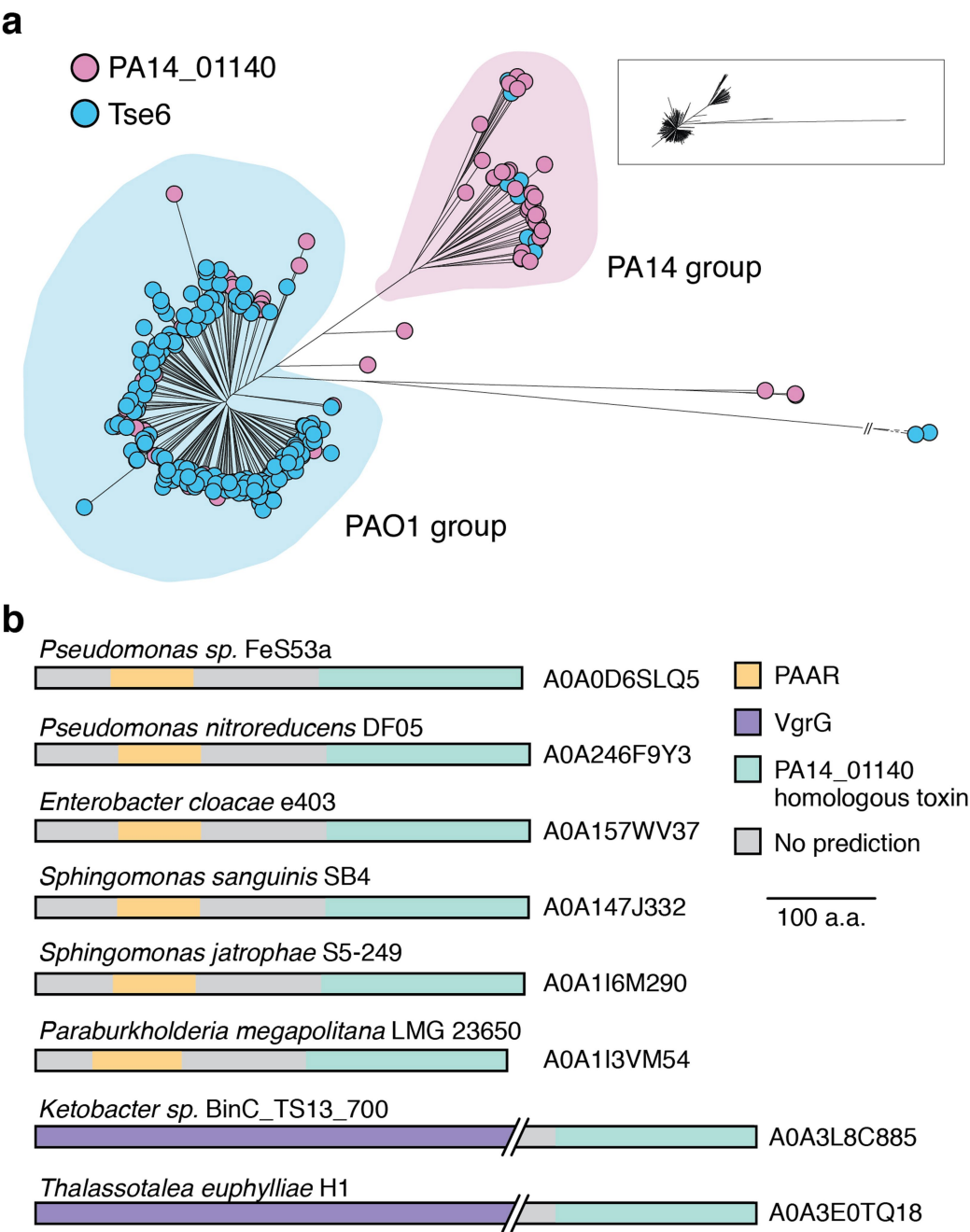
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1735-9>.

Correspondence and requests for materials should be addressed to M.T.L. or J.C.W.

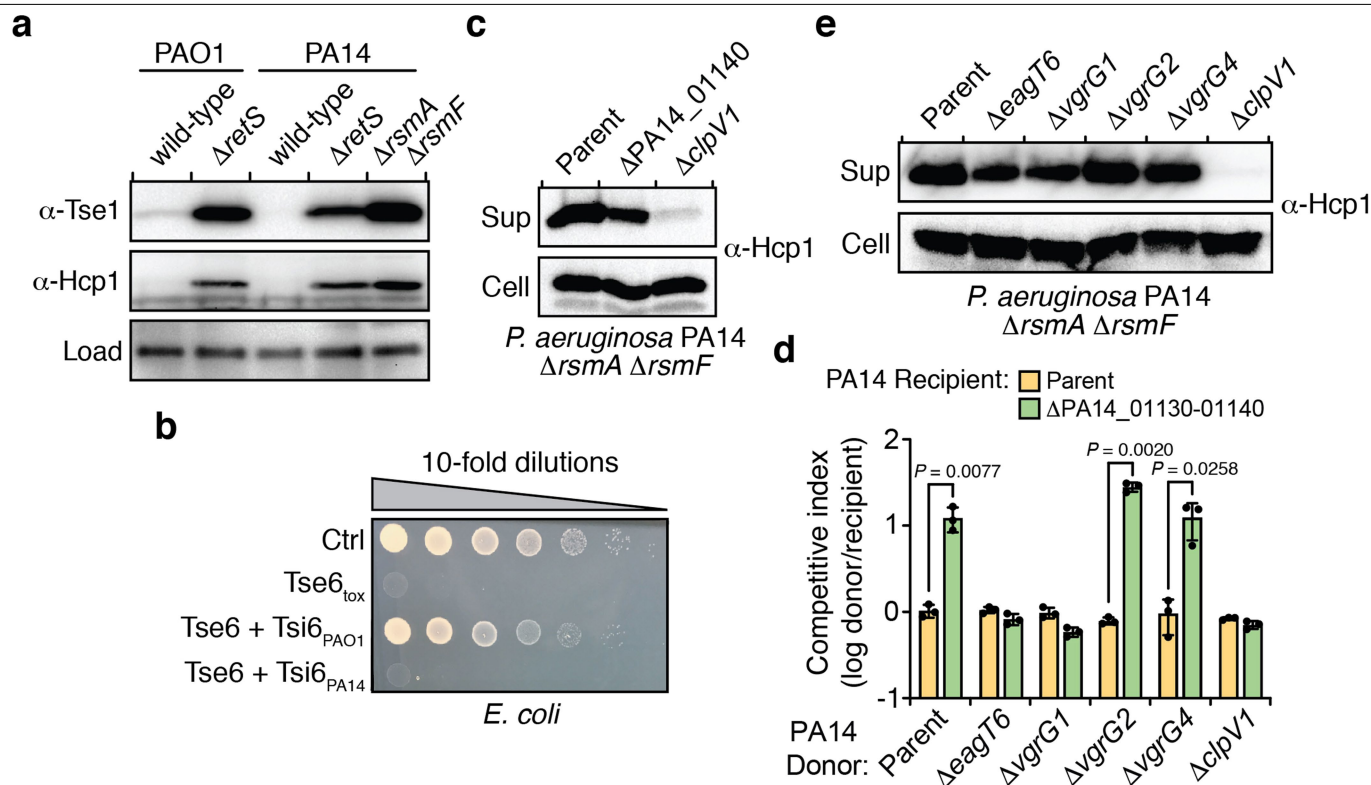
Peer review information *Nature* thanks Urs Jenal, Justin Nodwell and Jue Wang for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Homologues of PA14_01140 and Tse6 are enriched in *P. aeruginosa* PA14- and PAO1-related strains, respectively. **a**, Phylogenetic distribution of PA14_01140 (pink) and *tse6* (blue) within 326 *P. aeruginosa* genomes based on whole-genome single-nucleotide polymorphism (SNP) maximum likelihood analysis. Circles denote individual *P. aeruginosa* strains. Each clade is labelled according to its representative member. The miniaturized tree depicts true branch distance between each clade. The full

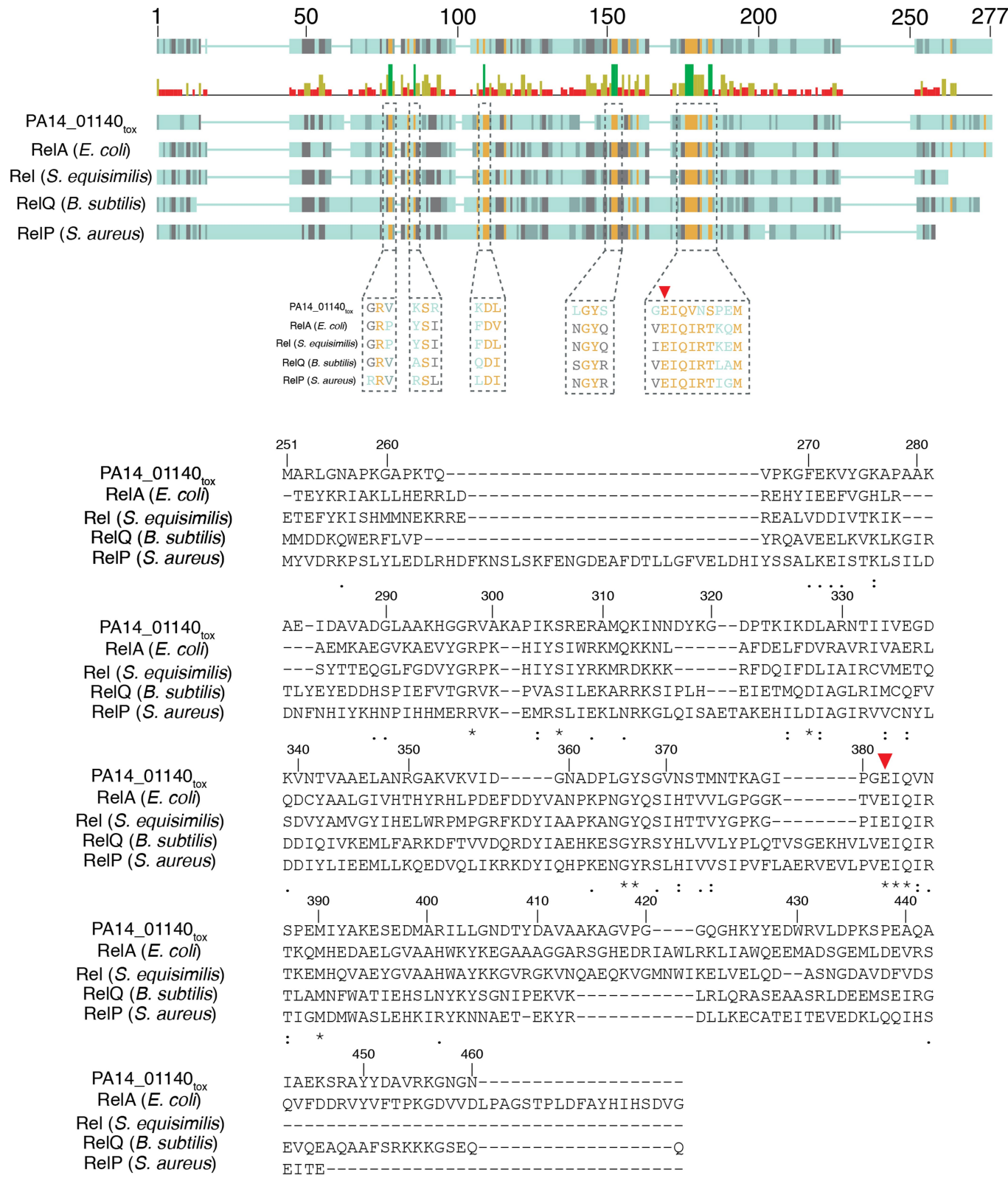
tree in Newick format, including bootstrap values, is provided as Supplementary Dataset 1. **b**, Proteins containing a domain homologous to the C terminus of PA14_01140 are found in several species of Proteobacteria. Homologues were identified using the HMMER webserver and candidate T6SS effectors were selected on the basis of the presence of predicted N-terminal domains known to facilitate export by the T6SS. The UniProt accession number for each identified protein is indicated.



Extended Data Fig. 2 | Characterization of the PA14_01140-PA14_01130-*tsi6* gene cluster.

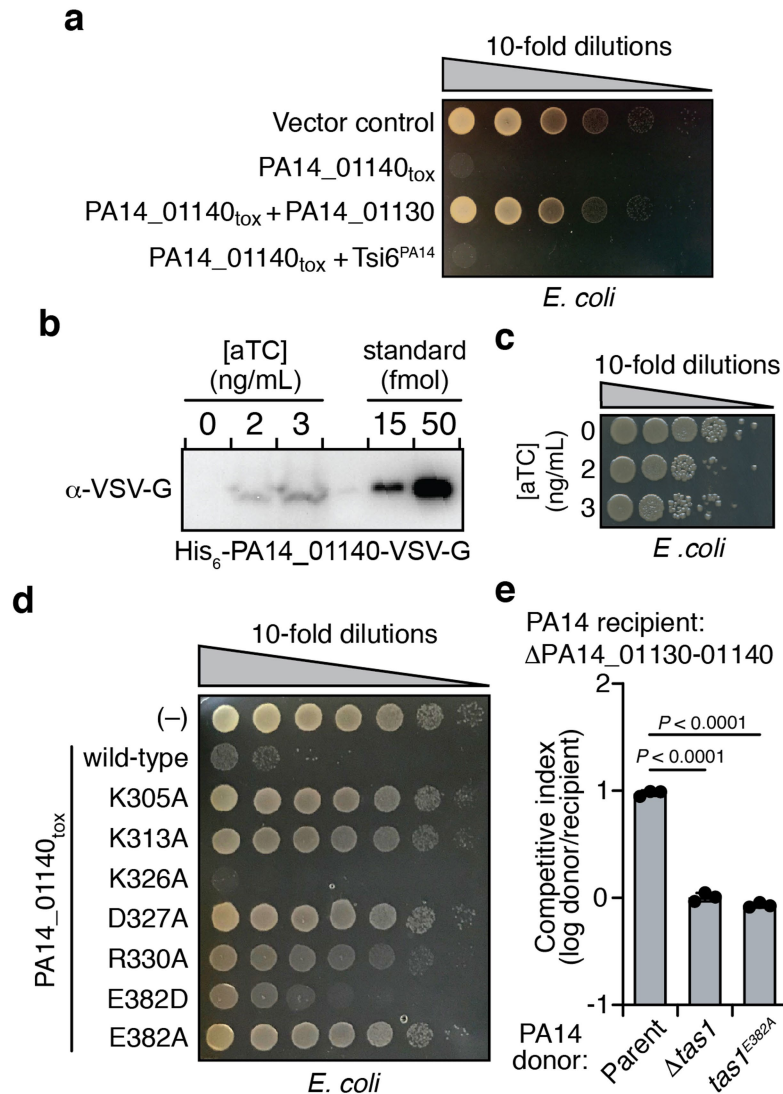
a, Expression of the conserved H1-T6SS effector Tse1 and the secreted H1-T6SS subunit Hcp1 are similar between *P. aeruginosa* PAO1 $\Delta retS$ and *P. aeruginosa* PA14 $\Delta rsmA \Delta rsmF$ by western blot analysis. A non-specific band that reacts with the anti-Tse1 antiserum was used as a loading control. **b**, Tsi6^{PA14} is not protective against Tse6-mediated intoxication. Viability of *E. coli* cells grown on solid medium harbouring inducible plasmids expressing Tse6_{tox}, Tse6_{tox} + Tsi6^{PAO1}, Tse6_{tox} + Tsi6^{PA14}, or an empty vector control. **c**, Mutational inactivation of PA14_01140 does not abrogate secretion of Hcp1. Western blot analysis of Hcp1 in the cell and supernatant (sup) fractions of the

indicated *P. aeruginosa* PA14 strains. **d**, Delivery of PA14_01140 into recipient cells requires the H1-T6SS exported protein VgrG1 and the Tse6-specific chaperone EagT6. Intraspecific growth competition assay between indicated PA14 donor and recipient strains. The parental strain genotype is $\Delta rsmA \Delta rsmF$. Mean \pm s.d. for $n = 3$ biological replicates; two-tailed, unpaired *t*-test. **e**, Mutational inactivation of *eagT6*, *vgrG1*, *vgrG2* and *vgrG4* does not abrogate H1-T6SS function. Western blot analysis of Hcp1 in the cell and supernatant fractions of the indicated *P. aeruginosa* PA14 strains. **a-c, e**, Data are representative of three independent experiments.



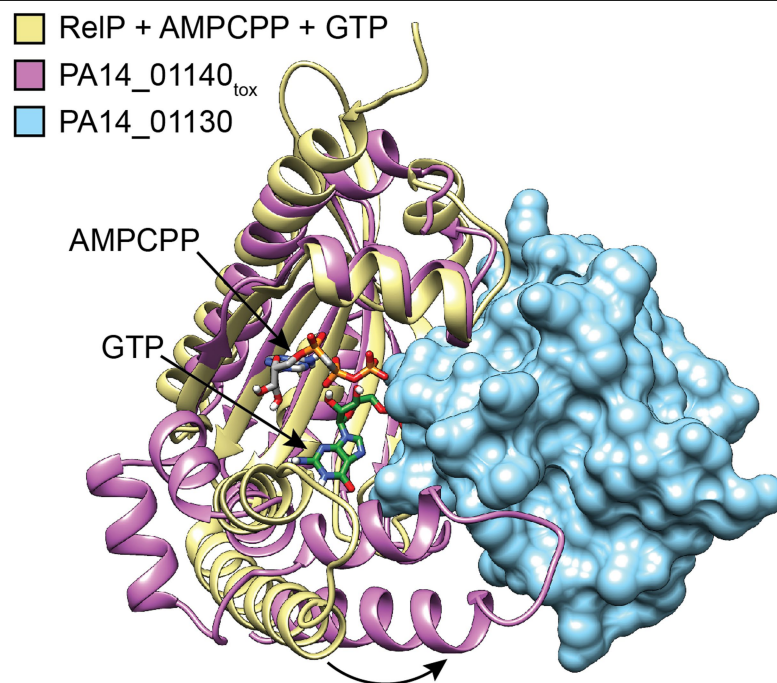
Extended Data Fig. 3 | PA14_01140_{tox} possesses remote homology to characterized (p)ppGpp synthetases. ClustalW alignment of PA14_01140_{tox}, the RSH domains of *E. coli* RelA and *Streptococcus equisimilis* Rel, and the small

alarmone synthetases RelQ and RelP from *B. subtilis* and *S. aureus*, respectively. Dashed boxes represent regions of high sequence homology. The catalytic glutamic acid is indicated by a red triangle.



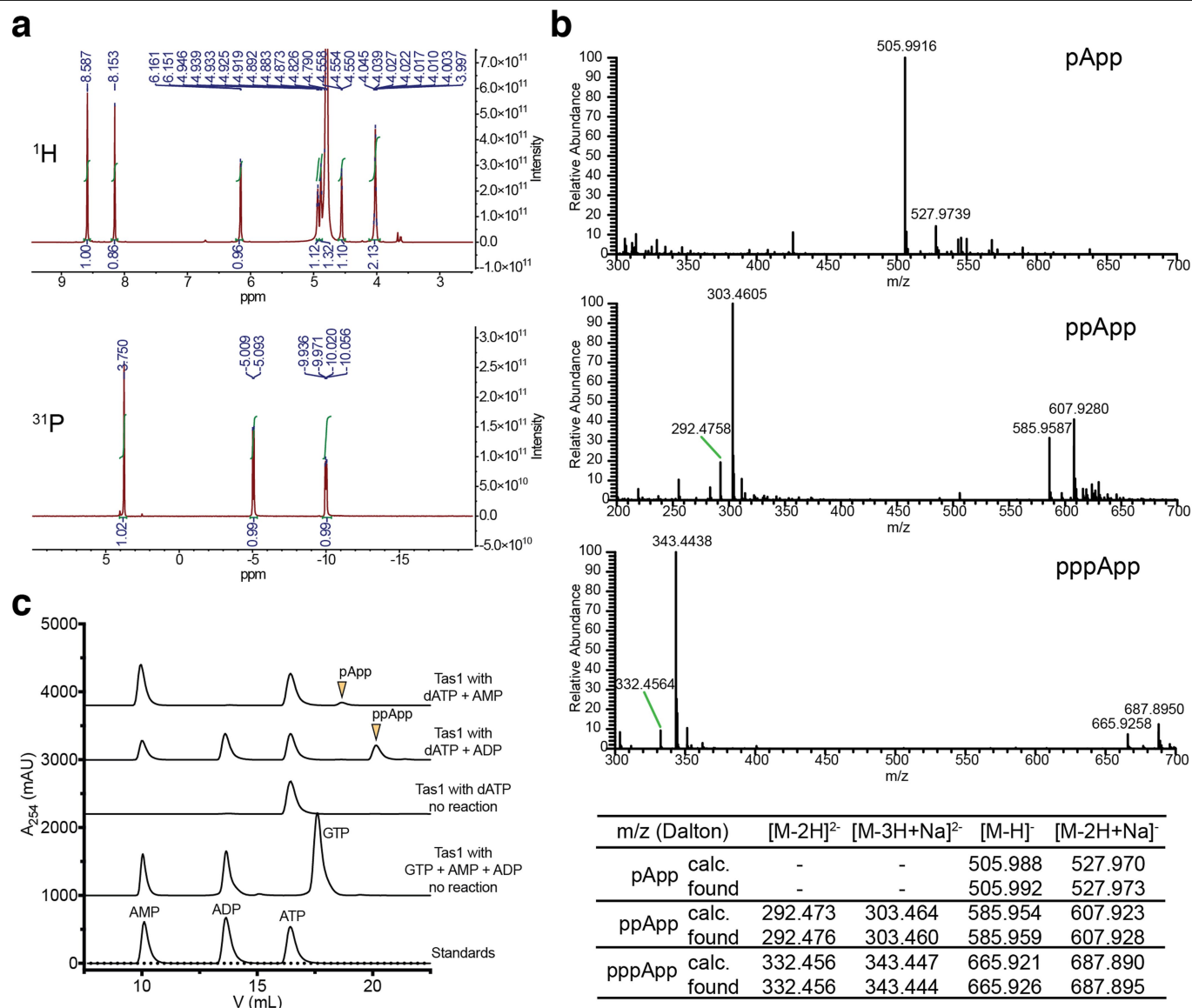
Extended Data Fig. 4 | The C-terminal domain of PA14_01140 (PA14_01140_{tox}) is toxic when expressed in *E. coli*. **a**, PA14_01130 but not Tsi6^{PA14} inhibits PA14_01140_{tox}-mediated toxicity. Viability of *E. coli* cells grown on solid medium harbouring inducible plasmids expressing PA14_01140_{tox}, PA14_01140_{tox} + PA14_01130, PA14_01140_{tox} + Tsi6^{PA14}, or an empty vector control. **b**, **c**, PA14_01140_{tox} is toxic to *E. coli*, even when expressed at approximately three copies per cell. **b**, Western blot analysis of pull-downs from *E. coli* expressing His₆-PA14_01140_{tox}-VSV-G in the presence of the indicated concentrations of aTC inducer (see Methods). **c**, Viability of *E. coli* cells expressing His₆-PA14_01140_{tox}-VSV-G in the presence of the indicated aTC concentrations for 15 min. **d**, Amino acid residues in PA14_01140_{tox} that structurally align with known pyrophosphate donor ATP-interacting residues

in RelQ are required for PA14_01140_{tox}-mediated toxicity. Viability of *E. coli* cells grown on solid medium harbouring inducible plasmids expressing PA14_01140_{tox}, each of the indicated PA14_01140_{tox} point mutants or an empty vector control. Lysine 326 is a residue located within the PA14_01140_{tox} active site that is not predicted to interact with the pyrophosphate donor ATP. **e**, Glutamate 382 is required for PA14_01140-based intoxication of susceptible recipient cells. Outcome of intraspecific growth competitions between the indicated PA14 donor strains and a ΔPA14_01130-1140 recipient. The parental PA14 strain genotype is Δ*rsmA*Δ*rsmF*. The competitive index is normalized to starting donor/recipient ratios. Mean ± s.d. for *n* = 3 biological replicates; two-tailed, unpaired *t*-test. **a–e**, Data are representative of two independent experiments.



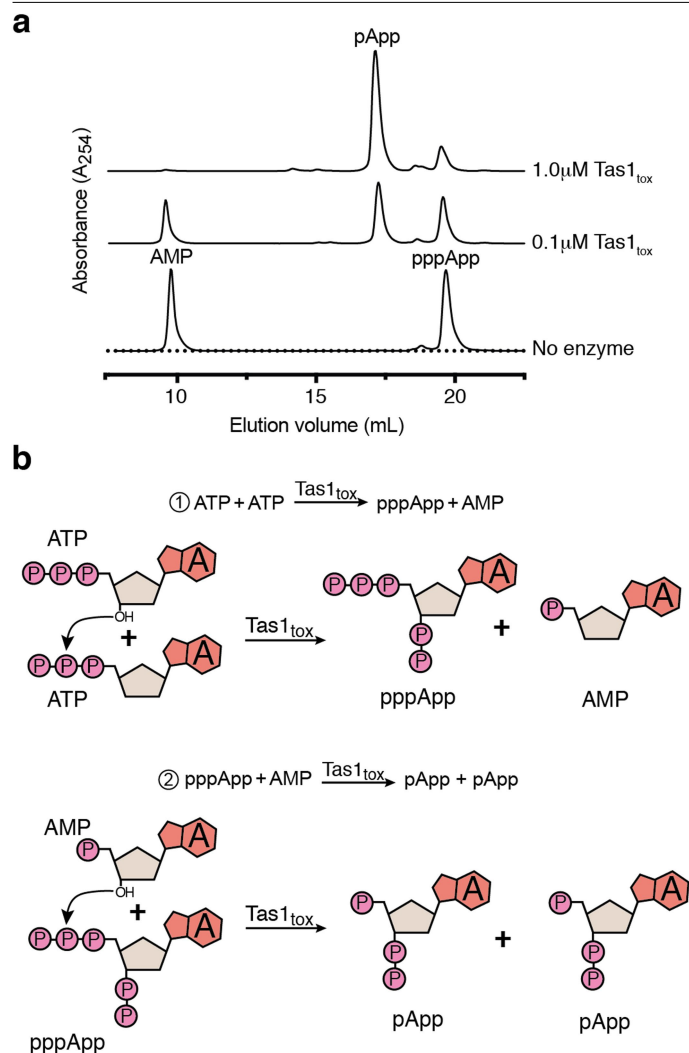
Extended Data Fig. 5 | Interaction with PA14_01130 distorts the predicted nucleotide acceptor site of PA14_01140_{tox}. Structural alignment between PA14_01140_{tox}-PA14_01130 complex and the (p)ppGpp synthetase RelP bound to the non-hydrolysable ATP analogue AMPCPP and a GTP acceptor nucleotide

(PDB code 6EWZ)⁴⁰. Two C-terminal α -helices of PA14_01140_{tox} that align with the GTP binding site of RelP are rotated approximately 30° as a consequence of their interaction with PA14_01130 (curved black arrow).

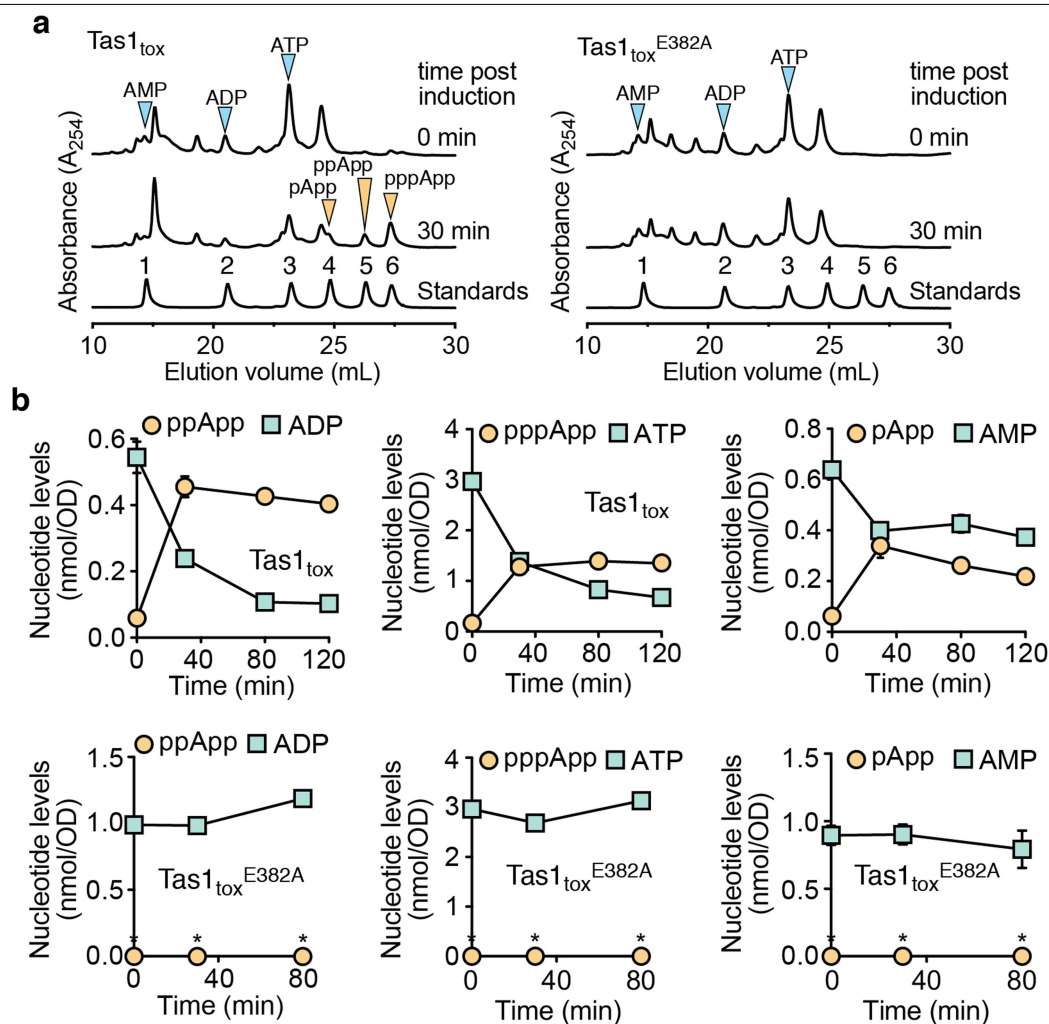


Extended Data Fig. 6 | Tas1 pyrophosphorylates the 3' hydroxyl group of adenosine nucleotides. a, ¹H (top) and ³¹P (bottom) NMR spectra of pApp. See Supplementary Table 2 for assignments. **b,** Negative mode electrospray mass spectra for pApp (top), ppApp (middle) and pppApp (bottom). Assignment of

major peaks is shown below the spectra. **c,** Anion-exchange traces of Tas1_{lox}-catalysed reactions with dATP or GTP as pyrophosphate donors. Arrowheads indicate 3' pyrophosphorylation products. **a–c,** Data are representative of two independent experiments.

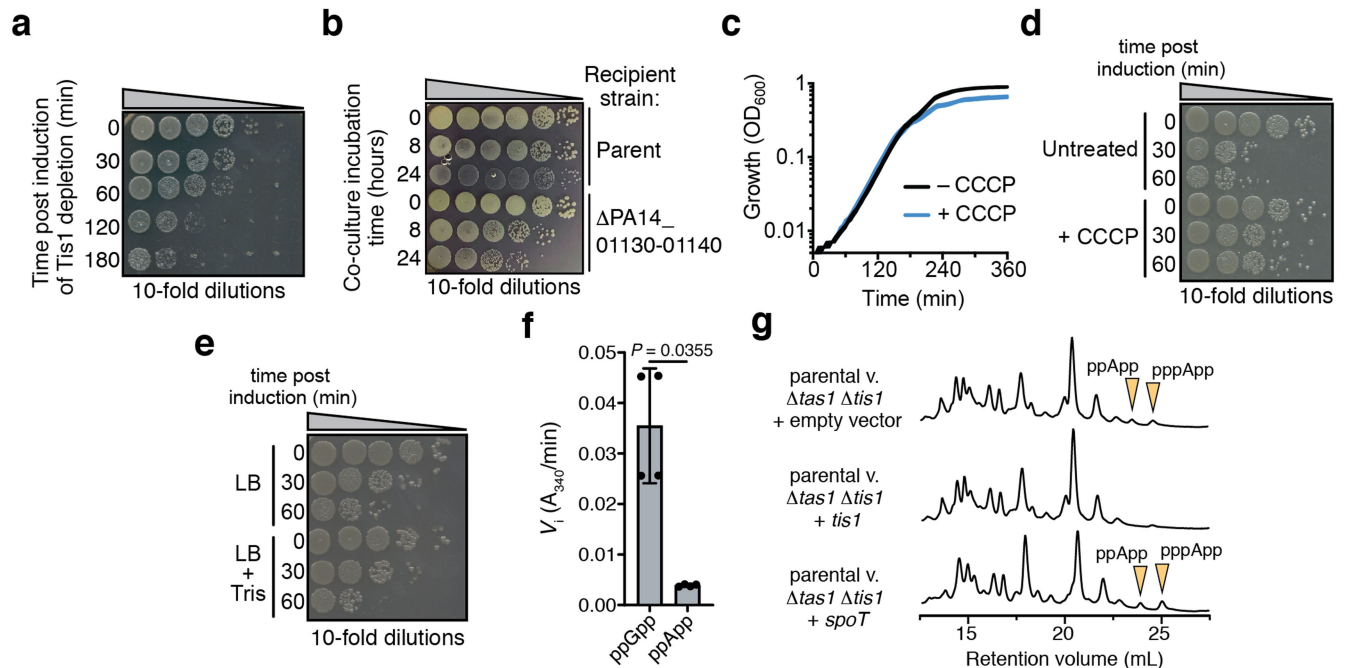


Extended Data Fig. 7 | Purified Tas1_{tox} can use pppApp as a pyrophosphate donor to pyrophosphorylate AMP and form pApp. a, Anion-exchange traces of pppApp and AMP after incubation with the indicated concentrations of Tas1_{tox} for 30 min at room temperature. A control lacking Tas1_{tox} is shown for comparison. Chromatogram is representative of two independent experiments. **b**, Mechanism of quantitative conversion of ATP to pApp. Only heteroatoms that participate in the reaction mechanism of pApp formation are shown.



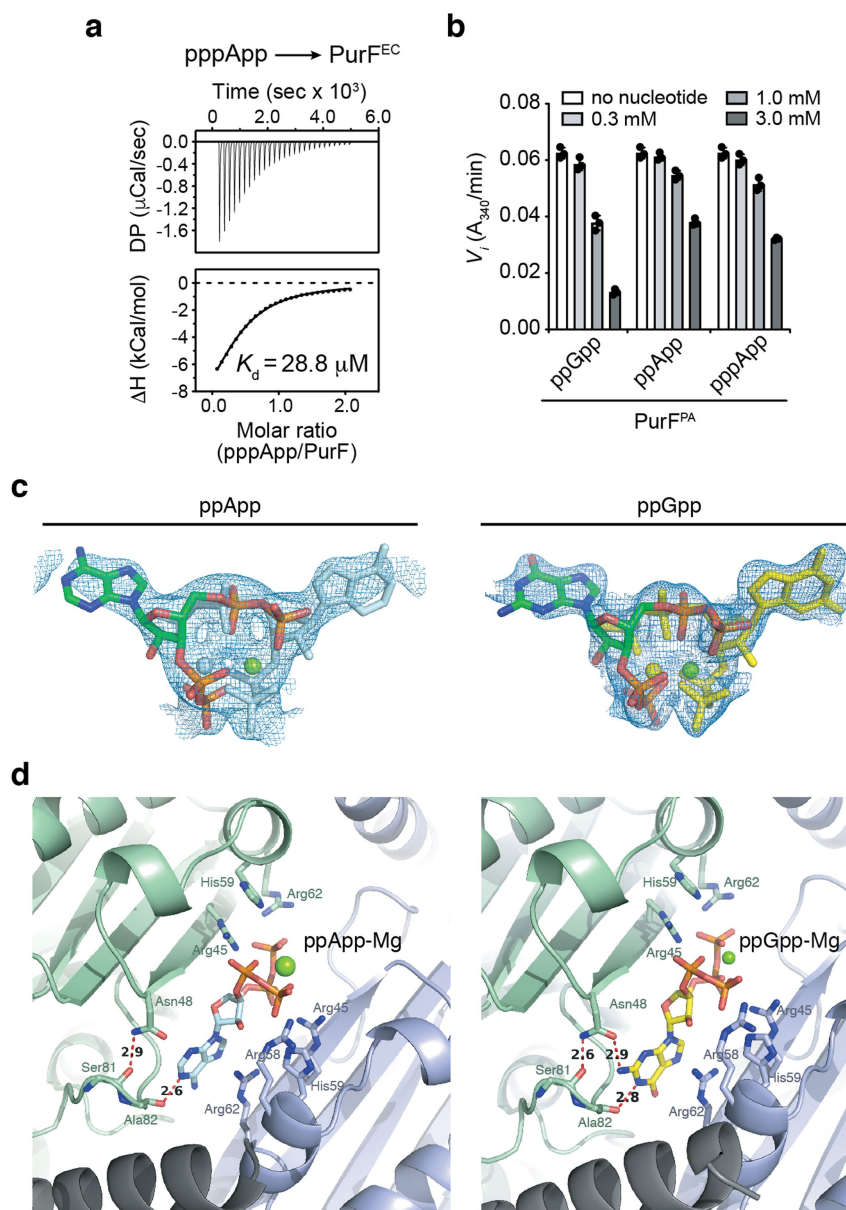
Extended Data Fig. 8 | Tas1_{tox} overexpression in *E. coli* leads to accumulation of (p)(p)App and a reduction in cellular 5' adenosine nucleotides. a, Anion-exchange chromatography traces of metabolites extracted from *E. coli* cells overexpressing Tas1_{tox} (left) or Tas1_{tox}^{E382A} (right) at the indicated time points. A trace generated from a mixture of standards containing an equimolar amount of AMP (1), ADP (2), ATP (3), pApp (4), ppApp (5) and pppApp (6) using the same gradient is shown for comparison. Peaks of adenosine

5'-nucleotides (AMP, ADP and ATP) and (p)(p)App are indicated by blue and orange arrowheads, respectively. Traces are representative of three independent experiments. **b**, Quantification of adenosine 5'-nucleotide and (p)(p)App levels in the *E. coli* strains from **a** as a function of time after induction. Mean \pm s.d. for metabolites extracted from $n = 3$ separate cultures. Asterisks indicate metabolites below the detection limit.



Extended Data Fig. 9 | The pmf-uncoupling ionophore CCCP but not the ppGpp-hydrolase domain of SpoT reduces the toxicity of TasI_{tox}. **a**, Tis1-depleted cells exhibited a reduction in viability over time. CFU plating of *P. aeruginosa* PA14 $\Delta retS \Delta sspB$ Tis1-D4 pPSV39::sspB cells at the indicated times after induction of SspB expression. **b**, TasI reduces the viability of susceptible recipient cells during interbacterial competition. CFU plating of the indicated *P. aeruginosa* PA14 recipient strains after co-culture with a parental donor strain at the indicated times. The parental PA14 strain genotype is $\Delta rsmA \Delta rsmF$. **c**, Steady-state growth of *E. coli* is not substantially affected by the presence of carbonyl cyanide *m*-chlorophenyl hydrazine (CCCP). Growth curves of *E. coli* cells harbouring the TasI_{tox} expression plasmid in LB medium with or without CCCP. Curves for $n = 3$ cultures are overlaid for each condition. **d**, TasI_{tox} toxicity is reduced in the presence of CCCP. Viability of *E. coli* cells following TasI_{tox} expression in the presence or absence of CCCP. Cells were

plated either before induction or at the indicated times after induction. **e**, Alkaline pH does not affect the ability of CCCP to reduce TasI_{tox}-dependent toxicity, indicating that the toxicity of TasI_{tox} is likely to arise from the generation of excessive membrane electrostatic potential. Cultures were untreated or conditioned to pH 8.0 using 25 mM Tris-HCl buffer immediately before induction. **f**, Activity of the ppGpp-hydrolase domain of SpoT against either ppGpp or ppApp. Initial velocities were normalized to hydrolase activity in the absence of either nucleotide. Mean \pm s.d. for enzymatic activity from $n = 4$ technical replicates; two-tailed, unpaired *t*-test. **g**, Anion-exchange chromatography traces of metabolites extracted from growth competition experiments between the indicated strains conducted on solid medium for 4 h. The parental strain is $\Delta rsmA \Delta rsmF$. Traces are representative of three independent experiments. **a**, **b**, **d**, **e**, Plates are representative of three independent experiments.



Extended Data Fig. 10 | (p)ppApp binds to and inhibits PurF in a similar manner to ppGpp. a, Isothermal calorimetry trace (top) and fitted isotherm (bottom) for the titration of 100 μM PurF^{EC} with 1 mM pppApp. Data are representative of two independent experiments. **b**, Changes to the activity of PurF^{PA} in the presence of indicated concentrations of ppGpp or (p)ppApp. Mean \pm s.d. for $n = 3$ separate reactions. **c**, $2F_o - F_c$ difference electron density maps of ppApp (left) and ppGpp (right, PDB code 6CZF) contoured at 0.4σ are shown in blue. Nucleotides are shown as stick models of two overlapping

ppApp-Mg²⁺ (coloured by heteroatom or light blue) or ppGpp-Mg²⁺ (coloured by heteroatom or yellow) complexes, related by a twofold rotational axis. **d**, Comparison of ppGpp and ppApp binding configuration within PurF^{EC}. The nucleotide-Mg²⁺ complexes are modelled at 0.5 occupancy because they lie on a crystallographic twofold rotational axis as shown in **c**. Relevant hydrogen bonding interactions and their distance in angstroms between PurF^{EC} residues and the purine rings of ppApp (left) or ppGpp (right) are shown with red dashed lines.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Bioinformatics: *Pseudomonas aeruginosa* genomes were downloaded from the *Pseudomonas* Genome Database (www.pseudomonas.com)
 Biochemical assays: Softmax Pro 6.22 (Molecular devices).
 LC-MS for metabolomics: XCalibur (Thermo Fisher Scientific).
 ITC: MicroCal Origin software (Malvern Panalytical).

Data analysis

Bioinformatics: open-reading frame prediction performed using Prodigal v2.6.1; identification of Tse6 and Tas1 orthologs completed using BLASTP v.2.8.1; determination of phylogenetic relationships between *P. aeruginosa* strains performed using PARSNP v.1.2 with PhiPack filtering; construction of phylogenetic tree was completed using RAXML-HP BlackBox v.8.2.10; visualization of phylogenetic tree performed using FigTree v.1.4.4.
 Anion-exchange chromatograms: integration performed with Unicorn 6.4 (GE Healthcare); regression and plotting performed using PRISM (Graphpad).
 LC-MS data for metabolomics: quantification of metabolites performed using XCalibur QuanBrowser 2.2 (Thermo Fisher Scientific).
 X-ray diffraction data: scaled, indexed and integrated using HKL3000 or XDS/XSCALE; molecular replacement was performed using PHASER; the model was refined using PHENIX with manual model building in COOT.
 Western blot quantification: image processing and densitometry analysis was completed using Fiji (distribution of ImageJ).
 Microscopy: Phase-contrast and propidium iodide-fluorescence images were processed using Metamorph software (Universal Imaging, PA).
 ITC: integrated and fitted using MicroCal Origin data analysis module for VP-ITC (Malvern Panalytical).
 Structures visualization: PyMOL 2.2.1 (Schrodinger LLC), UCSF Chimera 1.13.7 and UCSF ChimeraX 0.9.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Structural data for the Tas1-Tis1 complex (Fig. 2a-c and Extended Data Fig. 5) and the PurF-ppApp complex (Fig. 4e, f and Extended Data Fig. 10) has been deposited in Protein Data Bank (PDB) with the accession codes: 6OX6 and 6OTT, respectively. Raw metabolomics LC-MS data as sources of Figures 3b, 3g and 4a are available in Extended Data Table 2 and Supplementary Dataset 2. The full tree used to generate Extended Data Figure 1 is available in Newick format as Supplementary Dataset 1. 1H and 31P NMR assignments used for determining the structure of pApp are available in Extended Data Table 1. All other data generated or analyzed during this study are included in the manuscript (and its supplementary information files) or are available from the corresponding authors on request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For bacterial competition assays, enzymatic assays or metabolomics experiments, a sample size of 3 was chosen due to the significant and consistent differences between groups. Identical sample sizes have been used for similar methods that have been published previously.
Data exclusions	We did not observe outliers in our data that needed to be excluded.
Replication	Bacterial competition assays were completed in two independent experiments and the reported data error accounts for differences from three separate cultures per experiment. Metabolite analysis is performed with three independent bacterial cultures. Enzymatic assays are replicated with the same batch of enzyme and reagent stocks through independent pipetting. All attempts to replicate data were successful.
Randomization	This study does not involve subjects that require randomization.
Blinding	This study does not involve procedures that require blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Primary antibodies:
Hcp1: Custom polyclonal, GenScript, raised in rabbits, used at 1:3000 dilution. First described in Mougous et al., 2006, Science.
Tse1: Custom polyclonal, GenScript, raised in rabbits, used at 1:2000 dilution. First described in Russell et al., 2011, Nature.
VSV-G: Polyclonal (supplier: Sigma Aldrich (<https://www.sigmaaldrich.com/catalog/product/sigma/v4888?lang=en®ion=CA>)). Product number: V4888-200UG. Lot number: 058M4781V. Used at 1:5000 dilution.
Secondary antibody: Anti-rabbit IgG, HRP-linked antibody (supplier: Cell Signaling Technology). Product number: 7074S. Lot number: 28. Used at 1:5000 dilution.

Hcp1 was validated by western blot comparing a *Pseudomonas aeruginosa* strain expressing Hcp1 to a *P. aeruginosa* strain lacking the hcp1 gene.

Tse1 was validated by western blot comparing a *Pseudomonas aeruginosa* strain expressing Tse1 to a *P. aeruginosa* strain lacking the tse1 gene.

VSV-G was validated by western blot comparing a VSV-G-tagged to an untagged protein construct.

Validation for anti-rabbit: <https://media.cellsignal.com/coa/7074/28/7074-lot-28-coa.pdf>.

Activity of caspase-8 determines plasticity between cell death pathways

<https://doi.org/10.1038/s41586-019-1752-8>

Received: 12 February 2019

Accepted: 26 September 2019

Published online: 13 November 2019

Kim Newton^{1*}, Katherine E. Wickliffe¹, Allie Maltzman¹, Debra L. Dugger¹, Rohit Reja², Yue Zhang², Merone Roose-Girma³, Zora Modrusan³, Meredith S. Sagolla⁴, Joshua D. Webster⁴ & Vishva M. Dixit^{1*}

Caspase-8 is a protease with both pro-death and pro-survival functions: it mediates apoptosis induced by death receptors such as TNFR1¹, and suppresses necroptosis mediated by the kinase RIPK3 and the pseudokinase MLKL^{2–4}. Mice that lack caspase-8 display MLKL-dependent embryonic lethality⁴, as do mice that express catalytically inactive CASP8(C362A)⁵. *Casp8^{C362A/C362A} Mlkl^{-/-}* mice die during the perinatal period⁵, whereas *Casp8^{-/-} Mlkl^{-/-}* mice are viable⁴, which indicates that inactive caspase-8 also has a pro-death scaffolding function. Here we show that mutant CASP8(C362A) induces the formation of ASC (also known as PYCARD) specks, and caspase-1-dependent cleavage of GSDMD and caspases 3 and 7 in MLKL-deficient mouse intestines around embryonic day 18. Caspase-1 and its adaptor ASC contributed to the perinatal lethal phenotype because a number of *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-}* and *Casp8^{C362A/C362A} Mlkl^{-/-} Asc^{-/-}* mice survived beyond weaning. Transfection studies suggest that inactive caspase-8 adopts a distinct conformation to active caspase-8, enabling its prodomain to engage ASC. Upregulation of the lipopolysaccharide sensor caspase-11 in the intestines of both *Casp8^{C362A/C362A} Mlkl^{-/-}* and *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-}* mice also contributed to lethality because *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-} Casp11^{-/-}* (*Casp11* is also known as *Casp4*) neonates survived more often than *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-}* neonates. Finally, *Casp8^{C362A/C362A} Ripk3^{-/-} Casp1^{-/-} Casp11^{-/-}* mice survived longer than *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-} Casp11^{-/-}* mice, indicating that a necroptosis-independent function of RIPK3 also contributes to lethality. Thus, unanticipated plasticity in death pathways is revealed when caspase-8-dependent apoptosis and MLKL-dependent necroptosis are inhibited.

The caspase-8 scaffold mediates the production of cytokines by cancer cells that are exposed to TRAIL (also known as TNFSF10), recruiting RIPK1 via FADD to trigger NF-κB-dependent gene transcription^{6,7}. Caspase-8 has also been implicated in the activation of the NLRP3 inflammasome by Toll-like receptor 3 (TLR3)⁸. We tested whether these mechanisms contributed to intestinal atrophy and perinatal lethality in *Casp8^{C362A/C362A} Mlkl^{-/-}* mice⁵, as sterile inflammation was detected in the intestine as early as embryonic day 16.5 (E16.5). Leukocytes infiltrated the lumen of the intestine (Fig. 1a) and expression of genes such as *Cxcl10*, *Mgl2* and *Spic* was increased in comparison to *Mlkl^{-/-}* intestines (Fig. 1b). Intestines of *Casp8^{-/-} Mlkl^{-/-}* mice showed an intermediate level of gene expression, consistent with caspase-8 deficiency increasing the expression of pro-inflammatory genes during development⁹.

At E17.5 and E18.5, *Casp8^{C362A/C362A} Mlkl^{-/-}* embryos had higher levels of several serum cytokines and chemokines compared with *Mlkl^{-/-}* embryos (Extended Data Fig. 1a, b), whereas levels of cytokines and chemokines in *Casp8^{-/-} Mlkl^{-/-}* embryos were comparable to those in *Mlkl^{-/-}* or wild-type embryos (Extended Data Fig. 1b, c). Intestines

of some of the *Casp8^{C362A/C362A} Mlkl^{-/-}* mice also contained elevated cytokines and chemokines compared to intestines of *Mlkl^{-/-}* mice (Extended Data Fig. 1d). Villus atrophy was detected in two out of five *Casp8^{C362A/C362A} Mlkl^{-/-}* embryos at E17.5 (Extended Data Fig. 1e), indicating that the atrophy that was consistently observed⁵ at E18.5 has an acute onset between E17.5 and E18.5. The expression of CASP8(C362A) in intestines of *Mlkl^{-/-}* mice was comparable to that of wild-type caspase-8 in intestines of *Mlkl^{-/-}* mice (Extended Data Fig. 1f). Aberrant autophosphorylation of both RIPK1 and RIPK3 was observed in the intestines of *Casp8^{C362A/C362A} Mlkl^{-/-}* mice, consistent with the catalytic activity of caspase-8 suppressing necroptosis at the level of RIPK1⁵. RIPK3 autophosphorylation was also detected in the skin and liver of E18.5 *Casp8^{C362A/C362A} Mlkl^{-/-}* mice (Extended Data Fig. 1f).

NLRP3 deficiency did not prevent *Casp8^{C362A/C362A} Mlkl^{-/-}* mice from dying at birth (Table 1). *Casp8^{C362A/C362A} Mlkl^{-/-} Ripk1^{-/-}* mice, and—more rarely—*Casp8^{C362A/C362A} Mlkl^{-/-} Fadd^{-/-}* mice, were identified at 4 to 7 days after birth, but they were not found at weaning at 3 weeks of age (Table 1). *Mlkl^{-/-} Ripk1^{-/-}* littermates died before weaning, as expected,

¹Department of Physiological Chemistry, Genentech, South San Francisco, CA, USA. ²Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA, USA.

³Department of Molecular Biology, Genentech, South San Francisco, CA, USA. ⁴Department of Pathology, Genentech, South San Francisco, CA, USA. *e-mail: knewton@gene.com;

dixit@gene.com

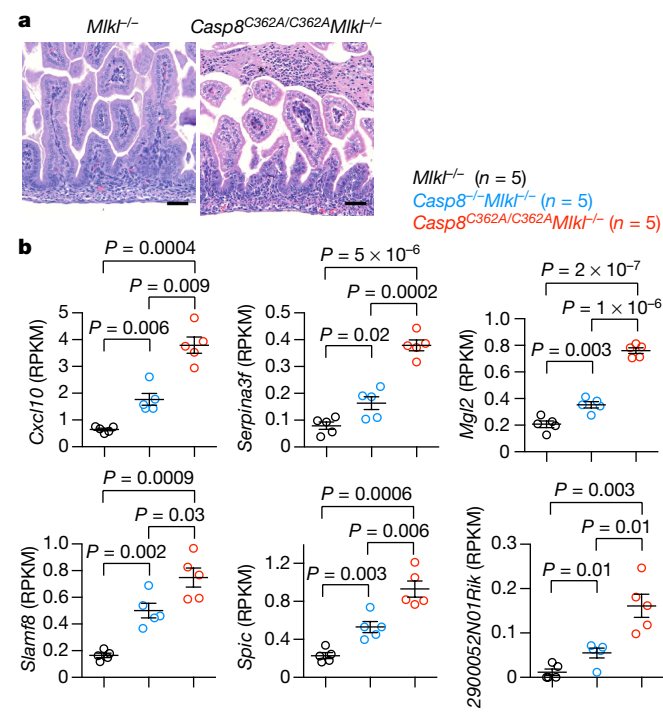


Fig. 1 | CASP8(C362A) causes MLKL-independent inflammation. **a**, E16.5 intestines. Infiltrating leukocytes are highlighted with an asterisk. Results are representative of two *Mkl1*^{-/-} and two *Casp8*^{C362A/C362A} *Mkl1*^{-/-} mice. Scale bars, 50 μm. **b**, Gene expression in E16.5 intestines analysed by RNA sequencing. RPKM, reads per kilobase per million reads. n = 5 per genotype. Circles, individual mice. Data are mean ± s.e.m. P values calculated by unpaired, two-tailed t-test with Welch's correction.

because RIPK1 suppresses the activation of caspase-8 and apoptosis in several tissues¹⁰. CASP8(C362A) cannot induce apoptosis⁵ and therefore we expected that *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Ripk1*^{-/-} mice were viable, similar to *Casp8*^{-/-} *Ripk3*^{-/-} *Ripk1*^{-/-} mice^{10–12}. Instead, our data suggest that inactive caspase-8 has a NLRP3-, RIPK1- and FADD-independent death-inducing scaffold function in the perinatal period.

Given that the concentrations of IL-18 and IL-1β, which are processed by caspase-1¹³, were elevated in *Casp8*^{C362A/C362A} *Mkl1*^{-/-} embryos, we tested whether pyroptotic cell death driven by caspase-1 or caspase-11 contributed to the death of *Casp8*^{C362A/C362A} *Mkl1*^{-/-} pups. Loss of caspase-1,

Table 1 | Numbers of P4–P7 offspring from intercrossing *Casp8*^{C362A/+} mice

Genetic background	<i>Casp8</i> ^{+/+}	<i>Casp8</i> ^{C362A/+}	<i>Casp8</i> ^{C362A/C362A}
<i>Mkl1</i> ^{-/-} <i>Nlrp3</i> ^{-/-}	36 ± 2 ^a	81 ± 2 ^a	0
<i>Mkl1</i> ^{-/-} <i>Fadd</i> ^{-/-}	22 ± 4 ^a	38 ± 8 ^a	3 ^a
<i>Mkl1</i> ^{-/-} <i>Ripk1</i> ^{-/-}	3 ± 4 ^a	20 ^a	18 ^a
<i>Mkl1</i> ^{-/-} <i>Casp11</i> ^{-/-}	36	103 ± 2 ^a	0
<i>Mkl1</i> ^{-/-} <i>Casp1</i> ^{-/-}	34 ± 1 ^a	87 ± 4 ^a	15 ± 22 ^a
<i>Mkl1</i> ^{-/-} <i>Casp1</i> ^{-/-} <i>Casp11</i> ^{-/-}	96 ± 13 ^a	197 ± 9 ^a	65 ± 21 ^a
<i>Mkl1</i> ^{-/-} <i>Asc</i> ^{-/-}	33 ± 1 ^a	61	11 ± 6 ^a
<i>Mkl1</i> ^{-/-} <i>Gsdmd</i> ^{-/-}	73	148 ± 8 ^a	2 ^a
<i>Ripk3</i> ^{-/-}	156 ± 7 ^a	355 ± 21 ^a	21 ± 7 ^a
<i>Ripk3</i> ^{RHIM/RHIM}	70	122 ± 2 ^a	5 ^a
<i>Ripk3</i> ^{-/-} <i>Casp1</i> ^{-/-} <i>Casp11</i> ^{-/-}	63 ± 1 ^a	153 ± 8 ^a	61 ± 2 ^a

P4–P7, postnatal days 4–7.
^aOffspring not found at weaning are listed separately from those that survived weaning.

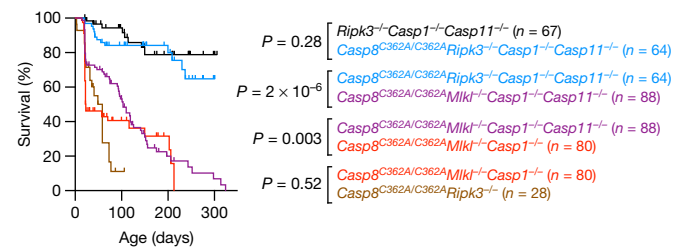


Fig. 2 | Caspase-1, caspase-11 and RIPK3 promote lethality in *Casp8*^{C362A/C362A} *Mkl1*^{-/-} mice. Kaplan–Meier curves of mouse survival. P values were calculated by two-sided Gehan–Breslow–Wilcoxon test. The number of mice differs from the list in Table 1, as some of the mice in the graph had a *Casp8*^{C362A/C362A} parent.

but not caspase-11, yielded 4- to 7-day-old *Casp8*^{C362A/C362A} *Mkl1*^{-/-} pups (Table 1), and 45% (36 out of 80) survived beyond weaning (Fig. 2). More pups survived past weaning if they also lacked caspase-11 (65 out of 88, or 74% of *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} pups). Thus, caspase-1 is a major driver of perinatal lethality in *Casp8*^{C362A/C362A} *Mkl1*^{-/-} mice, whereas caspase-11 seems to contribute to lethality around weaning. RIPK3 also had a non-necroptotic role, as more *Casp8*^{C362A/C362A} *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice survived and for longer than *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice (Fig. 2 and Table 1). Indeed, some *Casp8*^{C362A/C362A} *Ripk3*^{-/-} mice also survived to weaning (Table 1). Therefore, CASP8(C362A) interacts directly or indirectly with RIPK3, caspase-1 and caspase-11 to compromise mouse survival. Rare *Casp8*^{C362A/C362A} *Ripk3*^{RHIM/RHIM} pups expressing RIPK3 with a mutant RIP homotypic interaction motif (RHIM)¹⁴ were identified at 4–7 days but none survived to weaning (Table 1). Thus, RIPK3 may have a RHIM-independent function. RIPK3 co-immunoprecipitated with CASP8(C362A) from the intestines of E16.5 mice (Extended Data Fig. 1g), whereas caspases 1 and 11 were not detected in CASP8(C362A) complexes.

Casp8^{C362A/C362A} *Ripk3*^{-/-} mice were severely undersized (Extended Data Fig. 2a) and had anaemia (Extended Data Fig. 2b), splenomegaly (Extended Data Fig. 2c) and mild immune cell infiltrates in tissues such as the lungs (Extended Data Fig. 2d). *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} mice had a similar phenotype, but were less stunted (Extended Data Fig. 2). Expansion of the red pulp by extramedullary haematopoiesis contributed to the splenomegaly (Extended Data Fig. 2d). *Casp8*^{C362A/C362A} *Ripk3*^{-/-}, *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} and *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice did not have an expanded B220⁺ CD3⁺ T cell population as was found in *Casp8*^{-/-} *Mkl1*^{-/-} mice⁴ (Extended Data Figs. 3, 4a), although such cells were evident in the mesenteric lymph node of *Casp8*^{C362A/C362A} *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice (Extended Data Fig. 4a). Spleens of *Casp8*^{C362A/C362A} *Ripk3*^{-/-}, *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} and *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice contained fewer B220⁺ cells than spleens from littermate controls (Extended Data Fig. 3), consistent with a severe block in B cell differentiation in the bone marrow (Extended Data Fig. 4b, c). Bone marrow from *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} and *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice was also deficient in Lineage⁻ SCA-1⁺ KIT⁺ progenitor cells (Extended Data Fig. 4b, c). Deficits in B cells and Lineage⁻ SCA-1⁺ KIT⁺ cells were less severe in *Casp8*^{C362A/C362A} *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} mice. Finally, more *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-}, *Casp8*^{C362A/C362A} *Mkl1*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} and *Casp8*^{C362A/C362A} *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} CD4⁺ or CD8⁺ T cells had a CD62L^{low} CD44^{high} activated phenotype than their control counterparts (Extended Data Fig. 3). Therefore, CASP8(C362A) elicits several haematological abnormalities independent of RIPK3, MLKL, caspase-1 or caspase-11, although their presence appears to exacerbate the reduction in B cells and Lineage⁻ SCA-1⁺ KIT⁺ cells.

Pro-inflammatory signals increase expression of caspase-11¹⁵, and caspase-11 was more abundant in the intestines of *Casp8*^{C362A/C362A} *Mkl1*^{-/-} or

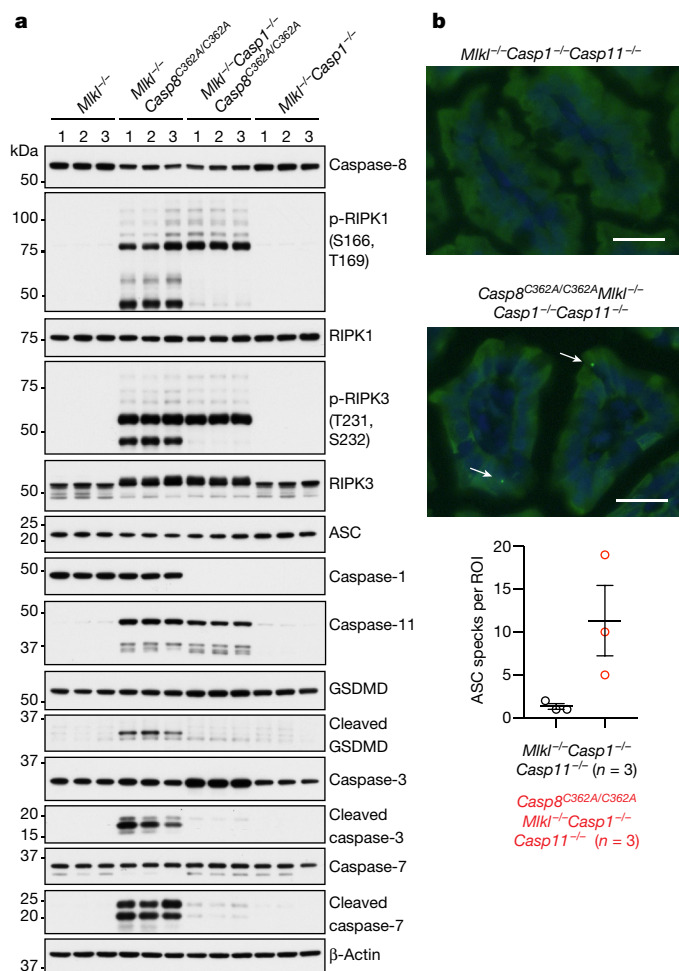


Fig. 3 | CASP8(C362A) promotes the formation of ASC specks. **a**, Western blots of E18.5 intestines. Lane numbers indicate different mice. $n = 3$ per genotype. Analysis of the β -actin loading control was performed after GSDMD. For gel source data, see Supplementary Fig. 1. **b**, Top, E18.5 intestines with ASC immunolabelling (green); DAPI was used to label the DNA (blue). Arrows, ASC specks. Scale bars, 25 μ m. Bottom, quantification of ASC specks. $n = 3$ per genotype. Circles, individual mice. Data are mean \pm s.e.m. $P = 0.13$, unpaired two-tailed t -test with Welch's correction.

Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-} mice compared with intestines of *Mkl1^{-/-}* or *Mkl1^{-/-} Casp1^{-/-}* mice (Fig. 3a). Single-cell RNA-sequencing analysis confirmed that a subset of *Casp8^{C362A/C362A} Mkl1^{-/-}* intestinal cells, in five out of eight different cell clusters, expressed higher levels of *Casp11* mRNA than their *Mkl1^{-/-}* counterparts (Extended Data Fig. 5). Given that caspase-11 is activated by lipopolysaccharide (LPS)¹⁶, we explored whether *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-}* pups had increased exposure to LPS from commensal bacteria owing to increased intestinal permeability. However, serum LPS levels were comparable in *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-}* and *Mkl1^{-/-} Casp1^{-/-}* 1-week-old littermates (Extended Data Fig. 6a) and the mice had comparable intestinal permeability at 3–4 weeks of age (Extended Data Fig. 6b). We speculate that increased caspase-11 levels in the gut of *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-}* mice lowers the threshold for activation of caspase-11 by LPS as the newborn gut is colonized, but this notion needs to be validated in a germ-free setting.

Loss of GSDMD (gasdermin D), a pore-forming protein that triggers cell lysis after cleavage by caspase-1¹⁶, did not allow *Casp8^{C362A/C362A} Mkl1^{-/-}* mice to survive to weaning (Table 1). Thus, cleavage of other caspase-1 substrates, such as caspase-7¹⁷, might be sufficient to induce lethality. Indeed, we detected caspase-1-dependent cleavage of caspase-7,

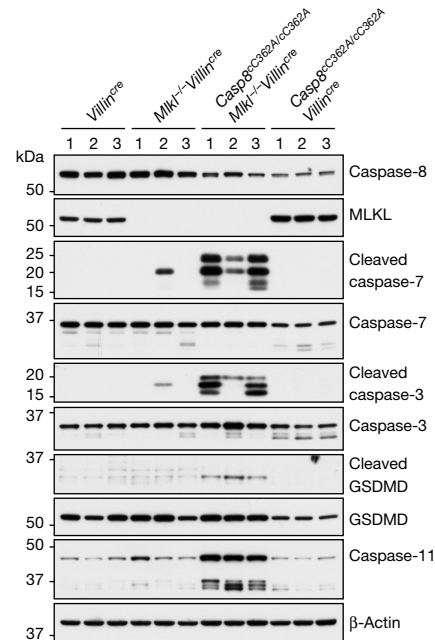


Fig. 4 | Expression of CASP8(C362A) in necroptosis-deficient intestinal epithelial cells triggers cleavage of caspase-1 substrates. Western blots of P0 intestines. Lane numbers indicate different embryos. $n = 3$ per genotype. Analysis of the β -actin loading control was performed after MLKL. For gel source data, see Supplementary Fig. 1.

caspase-3 and GSDMD in the intestines of *Casp8^{C362A/C362A} Mkl1^{-/-}* mice (Fig. 3a). Autophosphorylated forms of RIPK1 and RIPK3 also appeared to be cleaved in a caspase-1-dependent manner.

Caspase-8 interacts with the caspase-1 adaptor ASC in cells infected with *Francisella*¹⁸ or *Salmonella*¹⁹; we therefore hypothesized that CASP8(C362A) activates caspase-1 through ASC. Accordingly, ASC specks, which serve as a platform for the activation of caspase-1²⁰, were detected in the intestines of *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-} Casp11^{-/-}* mice but not in intestines of *Mkl1^{-/-} Casp1^{-/-} Casp11^{-/-}* mice (Fig. 3b). In transfection studies, both wild-type caspase-8 exposed to the pan-caspase inhibitor emricasan and CASP8(C362A) had a propensity to enter Triton X-100-insoluble aggregates, whereas wild-type caspase-8 that was not exposed to emricasan was detected only in the soluble fraction (Extended Data Fig. 6c). In addition, CASP8(C362A) and emricasan-exposed wild-type caspase-8 shifted co-transfected ASC into the insoluble fraction more than wild-type caspase-8 that was not exposed to emricasan. Consistent with these data, CASP8(C362A) co-localized with ASC in specks more often than wild-type caspase-8 (Extended Data Fig. 6d). Thus, inactive caspase-8 appears to nucleate ASC speck assembly for activation of caspase-1 in intestinal epithelial cells (IECs).

The pyrin domain of ASC interacts with the death effector domains (DEDs) in caspase-8²¹; we therefore tested whether the DED-containing pro-domain of caspase-8 was required for ASC to enter the insoluble fraction. Indeed, pro-domain residues 1–177 increased the amount of ASC in the insoluble fraction to the same extent as CASP8(C362A), whereas residues 212–481 carrying the C362A mutation behaved similar to wild-type caspase-8 (Extended Data Fig. 6e). Therefore, inactive caspase-8 may favour a conformation that unmasks the pro-domain for interactions with ASC. Accordingly, ASC deficiency yielded *Casp8^{C362A/C362A} Mkl1^{-/-}* survivors (Table 1), and limited cleavage of GSDMD, caspase-3 and caspase-7 in the intestine of *Casp8^{C362A/C362A} Mkl1^{-/-}* mice (Extended Data Fig. 6f).

It was notable that CASP8(C362A) was deleterious in the gut, but not in other tissues. Expression of ASC, caspase-1 and caspase-8 was not restricted to the intestine (Extended Data Fig. 1f). It was also unclear

Table 2 | Number of offspring from intercrossing *Casp8^{cc362A/+}* mice

Age	Background	Cre	<i>Casp8^{+/+}</i>	<i>Casp8^{cc362A/+}</i>	<i>Casp8^{cc362A/cc362A}</i>
P4–P7	Wild type	Vav	19	41	0
P4–P7	<i>Mlkl^{-/-}</i>	Vav	8	17	6
P4–P7	Wild type	Villin	12	32	12
P4–P7	<i>Mlkl^{-/-}</i>	Villin	48	77	2
P1	<i>Mlkl^{-/-}</i>	Villin	9	15	1
P0	<i>Mlkl^{-/-}</i>	Villin	13	18	14 + 3 ^a

^aThree mice were found dead at P0.

whether the gut phenotype contributed to perinatal lethality. Given that leukocytes infiltrated the lumen of the developing *Casp8^{cc362A/cc362A} Mlkl^{-/-}* intestine, we investigated whether restricting expression of CASP8(C362A) to either haematopoietic cells or IECs with a conditional *Casp8^{cc362A}* allele (Extended Data Fig. 7a) caused perinatal lethality. *Casp8^{cc362A/cc362A}* mice bearing a *Vav^{cre}* (also known as *Vav1*) transgene²² to express CASP8(C362A) in haematopoietic and endothelial cells were not born (Table 2), similar to *Casp8^{-/-}* and *Casp8^{cc362A/cc362A} Mlkl^{-/-}* mice^{1,5}. Lethality was due to aberrant necroptosis as *Casp8^{cc362A/cc362A} Mlkl^{-/-} Vav^{cre}* mice were viable (Table 2). *Casp8^{cc362A/cc362A} Mlkl^{-/-} Vav^{cre}* mice were not stunted (Extended Data Fig. 7b), but had similar haematological abnormalities to *Casp8^{cc362A/cc362A} Mlkl^{-/-} Casp1^{-/-}* mice, including splenomegaly (Extended Data Fig. 7c, d), a block in B cell differentiation (Extended Data Fig. 7d, e) and an activated T cell phenotype (Extended Data Fig. 7f). MAC-1⁺GR-1⁺ cells dominated the largest *Casp8^{cc362A/cc362A} Mlkl^{-/-} Vav^{cre}* spleen, and a relatively small population of B220⁺CD3⁺ T cells was detected in only one of the two other spleens (Extended Data Fig. 7g). Thus, B220⁺CD3⁺ T cells do not appear to drive splenomegaly in *Casp8^{cc362A/cc362A} Mlkl^{-/-} Vav^{cre}* mice.

Casp8^{cc362A/cc362A} mice bearing a *Villin^{cre}* (also known as *Vil1*) transgene²³ to express CASP8(C362A) in IECs were born (Table 2), but at 3–5 weeks of age were smaller than their *Villin^{cre}* siblings (Extended Data Fig. 8a). They had multi-focal lymphohistiocytic and neutrophilic enteritis plus crypt hyperplasia in the small intestine, and multi-focal lymphohistiocytic typhlocolitis in the large intestine (Extended Data Fig. 8b). Lesion severity varied with less-affected *Casp8^{cc362A/cc362A} Villin^{cre}* mice surviving at least 19 weeks (Extended Data Fig. 8c). Notably, most *Casp8^{cc362A/cc362A} Mlkl^{-/-} Villin^{cre}* mice did not survive to 4–7 days after birth (Table 2). Rare survivors showed stunted growth with crypt hyperplasia and villous atrophy (Extended Data Fig. 8d–f). Three out of four *Casp8^{cc362A/cc362A} Mlkl^{-/-} Villin^{cre}* newborn or 1-day-old pups exhibited similar histological changes (Extended Data Fig. 8g). Therefore, expression of CASP8(C362A) in IECs is sufficient for intestinal atrophy and perinatal lethality, but only if the cells are not eliminated by necroptosis. Consistent with CASP8(C362A) promoting cell death in the absence of MLKL, the intestines of newborn *Casp8^{cc362A/cc362A} Mlkl^{-/-} Villin^{cre}* mice exhibited increased cleavage of GSDMD, caspase-7 and caspase-3 compared with intestines of newborn *Mlkl^{-/-} Villin^{cre}* or *Casp8^{cc362A/cc362A} Villin^{cre}* mice (Fig. 4).

Collectively, our data reveal unexpected crosstalk between the apoptosis, necroptosis and pyroptosis cell death pathways. Previous studies have shown that active caspase-8 promotes activation of caspase-1 in response to bacterial infection^{24,25}. Our study suggests that caspase-1-dependent cell death also guards against inhibition of caspase-8, serving as a backup cell death mechanism when necroptosis is compromised. This crosstalk may have evolved as a defence against viruses that encode inhibitors of both caspase-8 and the protein interactions that

promote MLKL-dependent necroptosis²⁶. It is tempting to speculate that v-FLIPs expressed by several γ -herpesviruses and the poxvirus *Molluscum contagiosum* might induce such crosstalk by inhibiting necroptosis²⁷ and the activation of caspase-8²⁸.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1752-8>.

- Varfolomeev, E. E. et al. Targeted disruption of the mouse Caspase 8 gene ablates cell death induction by the TNF receptors, Fas/Apo1, and DR3 and is lethal prenatally. *Immunity* **9**, 267–276 (1998).
- Oberst, A. et al. Catalytic activity of the caspase-8-FLIP_L complex inhibits RIPK3-dependent necrosis. *Nature* **471**, 363–367 (2011).
- Kaiser, W. J. et al. RIP3 mediates the embryonic lethality of caspase-8-deficient mice. *Nature* **471**, 368–372 (2011).
- Alvarez-Diaz, S. et al. The pseudokinase MLKL and the kinase RIPK3 have distinct roles in autoimmune disease caused by loss of death-receptor-induced apoptosis. *Immunity* **45**, 513–526 (2016).
- Newton, K. et al. Cleavage of RIPK1 by caspase-8 is crucial for limiting apoptosis and necroptosis. *Nature* **574**, 428–431 (2019).
- Henry, C. M. & Martin, S. J. Caspase-8 acts in a non-enzymatic role as a scaffold for assembly of a pro-inflammatory “FADDosome” complex upon TRAIL stimulation. *Mol. Cell* **65**, 715–729 (2017).
- Hartwig, T. et al. The TRAIL-induced cancer secretome promotes a tumor-supportive immune microenvironment via CCR2. *Mol. Cell* **65**, 730–742 (2017).
- Kang, S. et al. Caspase-8 scaffolding function and MLKL regulate NLRP3 inflammasome activation downstream of TLR3. *Nat. Commun.* **6**, 7515 (2015).
- Kang, T. B., Jeong, J. S., Yang, S. H., Kovalenko, A. & Wallach, D. Caspase-8 deficiency in mouse embryos triggers chronic RIPK1-dependent activation of inflammatory genes, independently of RIPK3. *Cell Death Differ.* **25**, 1107–1117 (2018).
- Rickard, J. A. et al. RIPK1 regulates RIPK3–MLKL-driven systemic inflammation and emergency hematopoiesis. *Cell* **157**, 1175–1188 (2014).
- Dillon, C. P. et al. RIPK1 blocks early postnatal lethality mediated by caspase-8 and RIPK3. *Cell* **157**, 1189–1202 (2014).
- Kaiser, W. J. et al. RIP1 suppresses innate immune necrotic as well as apoptotic cell death during mammalian parturition. *Proc. Natl Acad. Sci. USA* **111**, 7753–7758 (2014).
- Afonina, I. S., Müller, C., Martin, S. J. & Beyaert, R. Proteolytic processing of interleukin-1 family cytokines: variations on a common theme. *Immunity* **42**, 991–1004 (2015).
- Newton, K. et al. RIPK1 inhibits ZBP1-driven necroptosis during development. *Nature* **540**, 129–133 (2016).
- Schauvliege, R., Vanrobaeys, J., Schotte, P. & Beyaert, R. Caspase-11 gene expression in response to lipopolysaccharide and interferon- γ requires nuclear factor- κ B and signal transducer and activator of transcription (STAT) 1. *J. Biol. Chem.* **277**, 41624–41630 (2002).
- Aglietti, R. A. & Dueber, E. C. Recent insights into the molecular mechanisms underlying pyroptosis and gasdermin family functions. *Trends Immunol.* **38**, 261–271 (2017).
- Lamkanfi, M. et al. Targeted peptidocentric proteomics reveals caspase-7 as a substrate of the caspase-1 inflammasomes. *Mol. Cell. Proteomics* **7**, 2350–2363 (2008).
- Pierini, R. et al. AIM2/ASC triggers caspase-8-dependent apoptosis in *Francisella*-infected caspase-1-deficient macrophages. *Cell Death Differ.* **19**, 1709–1721 (2012).
- Man, S. M. et al. *Salmonella* infection induces recruitment of caspase-8 to the inflammasome to modulate IL-1 β production. *J. Immunol.* **191**, 5239–5246 (2013).
- Stutz, A., Horvath, G. L., Monks, B. G. & Latz, E. ASC speck formation as a readout for inflammasome activation. *Methods Mol. Biol.* **1040**, 91–101 (2013).
- Vajihala, P. R. et al. The inflammasome adaptor ASC induces procaspase-8 death effector domain filaments. *J. Biol. Chem.* **290**, 29217–29230 (2015).
- Georgiades, P. et al. *Vav^{cre}* transgenic mice: a tool for mutagenesis in hematopoietic and endothelial lineages. *Genesis* **34**, 251–256 (2002).
- Madison, B. B. et al. *cis* elements of the Villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277**, 33275–33283 (2002).
- Weng, D. et al. Caspase-8 and RIP kinases regulate bacteria-induced innate immune responses and cell death. *Proc. Natl Acad. Sci. USA* **111**, 7391–7396 (2014).
- Philip, N. H. et al. Caspase-8 mediates caspase-1 processing and innate immune defense in response to bacterial blockade of NF- κ B and MAPK signaling. *Proc. Natl Acad. Sci. USA* **111**, 7385–7390 (2014).
- Nailwal, H. & Chan, F. K. Necroptosis in anti-viral inflammation. *Cell Death Differ.* **26**, 4–13 (2019).
- Chan, F. K. et al. A role for tumor necrosis factor receptor-2 and receptor-interacting protein in programmed necrosis and antiviral responses. *J. Biol. Chem.* **278**, 51613–51621 (2003).
- Thome, M. et al. Viral FLICE-inhibitory proteins (FLIPs) prevent apoptosis induced by death receptors. *Nature* **386**, 517–521 (1997).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Mice

All mouse studies complied with relevant ethics regulations and were approved by the Genentech Institutional Animal Care and Use Committee. *Vav^{cre}* mice²², *Villin^{cre}* mice²³, *Asc^{-/-}* mice²⁹, *Nlrp3^{-/-}* mice³⁰, *Casp1^{-/-}* *Casp11^{-/-}* mice³¹, *Casp11^{-/-}* mice³¹, *Casp1^{-/-}* mice³², *Gsdmd^{-/-}* mice³², *Ripk1^{+/-}* mice³³, *Ripk3^{-/-}* mice³³, *Ripk3^{RHIM/RHIM}* mice¹⁴, *Casp8^{+/-}* mice³³, *Mkl1^{-/-}* mice³⁴, *Casp8^{C362A/+}* mice⁵ and *Fadd^{+/-}* mice⁵ have previously been described. *Casp8^{C362A/+}* mice were generated at Genentech using C57BL/6 C2 embryonic stem cells (Extended Data Fig. 7a). Genotyping primers (5'-CGTAGAGCAGTCACAGATC-3', 5'-CCTGAGCGGTCCTTCT-3' and 5'-ATAGTGTCACCTAAATCGTATGT-3') amplified 266-bp wild-type, 371-bp *Casp8^{C362A}* and 300-bp *Casp8^{C362A}* DNA fragments. Alleles were maintained on a C57BL/6J (*Villin^{cre}*) or C57BL/6N (all others) genetic background.

For timed pregnancies, mice were designated E0.5 on the morning a vaginal plug was detected. Newborn mice were designated P0 on the day their birth was detected. Serum samples were analysed by Luminex assay using mouse cytokine/chemokine panels (Bio-Rad). In Extended Data Fig. 1b, serum IL-1 β was measured by enzyme-linked immunosorbent assay (ELISA; MSD, V-PLEX Mouse IL-1 β Kit K152QPD). Peripheral blood was analysed using an automated haematology analyser Sysmex XT-2000iV v.00-13. Serum LPS levels were measured by AssayGate. To measure intestinal permeability, mice aged 3–4 weeks were fasted for 4.5 h (with their regular ad libitum water supply) and then dosed with 0.5 mg fluorescein isothiocyanate (FITC)-conjugated dextran (average molecular mass 4 kDa, Sigma) per g body weight by oral gavage. Serum was collected 5 h later and serum FITC–dextran content was measured in a SpectraMax Gemini plate reader (Molecular Devices) with Softmax Pro 5.4.1.

Male and female mice ranging in age from E16.5 to P324 were analysed. Calculations were not performed to determine sample sizes. With the exception of ASC speck quantification in the intestine (Fig. 3b), analyses were not performed blinded to genotype. The experiments were not randomized.

Flow cytometry

Splenocytes, mesenteric lymph node and bone marrow cells were labelled with antibodies from BD Biosciences (anti-KIT–APC, 553356; anti-CD44–APC, 559250; anti-CD3–APC–Cy7, 557596; anti-GR-1–BV421, 562709; anti-IgM–BV421, 562595; anti-SCA-1–BV421, 562729; anti-B220–FITC, 553088; anti-B220–V500, 561226; anti-CD3–FITC, 553061; anti-CD4–FITC, 553651; anti-CD5–FITC, 553021; anti-CD8–FITC, 553031; anti-GR-1–FITC, 553127; anti-MAC-1–FITC, 553310; anti-TER-119–FITC, 557915; anti-CD4–PE, 553653; anti-MAC-1–PE, 553311; anti-CD62L–PE–Cy7, 560516) in the presence of 2% normal rat serum and 1 $\mu\text{g ml}^{-1}$ anti-CD16/CD32 antibody (BD Biosciences, 2.4G2, 553142). Leukocytes were identified by their forward scatter (FSC) and side scatter profiles. Dead cells that stained with 7-AAD (BD Biosciences), plus doublets, identified by their FSC-A versus FSC-W profiles, were excluded from analyses. Data were acquired using a BD FACSCantoII (BD Biosciences) cytometer and BD FACSDiva 8.0. Data were analysed with FlowJo 10.3.

Cell culture

HEK293T cells (ATCC CRL-3216; tested for mycoplasma contamination but not authenticated) were cultured in the high glucose version of Dulbecco's modified Eagle medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum, 2 mM glutamine, 10 mM HEPES (pH 7.2), 1 \times non-essential amino acids solution, 100 U ml⁻¹ penicillin and 100 $\mu\text{g ml}^{-1}$ streptomycin. Cells were transfected for 14 h with FuGENE HD (Promega) and plasmids encoding mouse caspase-8 and/or mouse ASC. Full-length caspase-8 and truncations were cloned into pCMV-3tag6 (Agilent Technologies) with an N-terminal 3 \times -Flag tag.

Full-length ASC was cloned into pCMV-3tag7 (Agilent Technologies) with an N-terminal 3 \times -Myc tag.

Western blotting and immunoprecipitation

Tissues or cells were lysed in 20 mM Tris-HCl pH 7.5, 135 mM NaCl, 1.5 mM MgCl₂, 1 mM EGTA, 1% Triton X-100, 10% glycerol, phosphatase inhibitor (Roche) and Halt protease inhibitor cocktail (Pierce). Tissues were mechanically disrupted with a bead mill homogenizer (Omni International) and insoluble material was removed by centrifugation at 20,000g before addition of LDS sample buffer. For HEK293T cells, lysates were centrifuged at 20,000g and the soluble fraction was removed. The 1% Triton X-100-insoluble pellet was washed once with lysis buffer before sonication in LDS sample buffer.

Western blot antibodies that recognized RIPK3 (1G6.1.4, Genentech; or NBPI-77299, Novus Biologicals), phosphorylated RIPK3 Thr231, Ser232 (GEN135-35-9, Genentech), RIPK1 (610459, BD Biosciences; 3493, Cell Signaling Technologies; or 10C7.3.1, Genentech), phosphorylated RIPK1 Ser166, Thr169 (GEN150-33-4, Genentech), MLKL (MABC604, EMD Millipore), β -actin (69100, MP Biomedicals), FADD (1.28E12, Genentech), caspase-8 (1G12, Enzo Life Sciences), FLIP (2.21H2, Genentech), ASC (8E4, Genentech), caspase-1 (4B4, Genentech), caspase-11 (17D9, Novus Biologicals), GSDMD (GN20-13, Genentech), cleaved GSDMD (50928, Cell Signaling Technology), caspase-7 (9492, Cell Signaling Technology), cleaved caspase-7 (9491, Cell Signaling Technology), caspase-3 (9662, Cell Signaling Technology), cleaved caspase-3 (9664, Cell Signaling Technology), Flag (A8592, Sigma) and Myc (GTX21261, Genetex). Caspase-8 was immunoprecipitated with anti-caspase-8 antibody (Abcam ab138485).

Immunofluorescence

The intestines of E18.5 embryos were fixed in 4% paraformaldehyde in PBS for 90 min at room temperature, washed three times in PBS and stored at 4 °C overnight in 30% sucrose in PBS before embedding in OCT. Frozen sections (10 μm) were blocked in PBS supplemented with 10% goat serum, 0.1% Triton X-100 and 0.1% saponin for 45 min at room temperature. Next, sections were incubated with 1:500 anti-ASC antibody (AdipoGen AG-25B-0006) for 1 h at room temperature, washed and incubated with 1:500 A488-conjugated anti-rabbit antibody (Life Technologies A11034) for 1 h at room temperature. Washed slides were mounted with Prolong Gold anti-fade reagent with DAPI (Invitrogen P36935). Images were acquired with a Leica DFC 365 FX camera and Leica Application Suite 4.6.0.

For ASC speck quantification, epifluorescence images were acquired on a Zeiss AxioImager M2 at 20 \times magnification. z-stacks were acquired at 0.5- μm intervals to capture the entire thickness of the tissue section. The resulting stacks were deconvolved using Huygens Professional 18.10 (Scientific Volume Imaging) to improve the signal-to-noise ratio and more easily visualize ASC specks. Deconvolved image stacks were imported into Slidebook 6.0 (Intelligent Imaging Innovations), processed into maximum-intensity projections and saved as TIFF images for quantification. Image intensity scaling was normalized across all images. Each image (450 $\mu\text{m} \times 330 \mu\text{m}$) was considered a single region of interest (ROI) and at least 5 ROIs were evaluated per animal. The maximum-intensity projection images were assessed blinded to genotype, and specks present in the small intestinal mucosa were manually counted in each ROI. Final quantification was reported as the mean number of ASC specks per ROI rounded to the nearest whole number.

HEK293T cells were plated in poly-D-lysine-coated 8-well chamber slides (75,000 cells per well) and then transfected with ASC and/or caspase-8 for 15 h. The cells were washed in PBS and fixed in 4% paraformaldehyde in PBS for 20 min at room temperature. Following three washes in PBS, the cells were blocked for 45 min in PBS containing 5% donkey serum, 5% goat serum, 1% Triton X-100 and 0.1% saponin. Immunolabelling was performed for 45 min at room temperature with 1:5,000 M2 anti-Flag antibody (Sigma, F3165) and 1:500 anti-ASC antibody

Article

(AdipoGen, AG25B-006). The cells were washed three times in PBS containing 0.1% Triton X-100 and then conjugated secondary antibodies were added at 1:500 dilution (A488 goat anti-mouse, Invitrogen, A11029; Cy5 donkey anti-rabbit, Jackson Laboratories, 711-176-152). After a 45-min incubation at room temperature, the cells were washed three times in PBS containing 0.1% Triton X-100 and mounted with Prolong Gold anti-fade reagent with DAPI.

Bulk RNA sequencing

Snap-frozen E16.5 intestines were thawed into RNAlater-ICE Frozen Tissue Transition Solution (ThermoFisher Scientific, AM7030) overnight at -20°C . RNA was extracted using a RNeasy Plus Mini Kit (QIAGEN, 74134). All samples had an RNA integrity number of ≥ 8.9 on an Agilent 2100 Bioanalyzer. RNA sample concentration was determined using a NanoDrop 8000 (Thermo Scientific) and the integrity of RNA was determined by Fragment Analyzer (Advanced Analytical Technologies). Libraries were prepared from 0.1 μg total RNA using a TruSeq Stranded Total RNA Library Prep Kit (Illumina). Library size was confirmed using 4200 TapeStation and High Sensitivity D1K screen tape (Agilent Technologies). Library concentration was determined using a Library quantification kit (KAPA). The libraries were multiplexed and sequenced on an Illumina HiSeq4000 (Illumina) to generate 30 million single-end 50-bp reads.

Raw FASTQ reads were aligned to the mouse reference genome (GRCm38/mm10) using GSNAP³⁵ (with parameters -M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1 --pairmax-rna = 200000 --clip-overlap). Reads were filtered to include only uniquely mapped reads. Differential expression analysis was performed using the voom/limma R package³⁶. Genes were considered differentially expressed if $\log_2(\text{gene expression in } Casp8^{C362A/C362A} Mkl^{-/-} / \text{gene expression in } Mkl^{-/-})$ was >1 or <-1 and $P < 0.05$ by two-sided, moderated t -test with Benjamini–Hochberg correction.

Single-cell RNA sequencing

E18.5 intestines were dissociated into single cells at 37°C for 30 min using a Worthington Biochemical Papain Dissociation Kit (LK003150). The density and viability of the single-cell suspensions were determined in a Vi-CELL XR cell counter (Beckman Coulter). All samples had $>90\%$ viable cells. Samples were processed using a Chromium Single Cell 3' Library and Gel bead kit v2 (10x Genomics). The cell density was used to impute the volume of single-cell suspension needed in the reverse transcription master mix, aiming to achieve around 6,000 cells per sample. cDNAs and libraries were prepared following the manufacturer's user guide (10x Genomics). Libraries were profiled by Bioanalyzer High Sensitivity DNA kit (Agilent Technologies) and quantified using Kapa Library Quantification Kit (Kapa Biosystems). Each library was sequenced in one lane of a HiSeq4000 (Illumina).

A gene-barcode matrix was generated for each sample using the 10x Genomics Cell Ranger pipeline v.3.0.2 (alignment to GRCm38 annotation, barcode assignment and unique molecular identifier counting). We detected 1,478 cells from the $Mkl^{-/-}$ intestine and 1,014 cells from the $Casp8^{C362A/C362A} Mkl^{-/-}$ intestine. Samples were filtered to remove empty droplets, possible doublets or multiplets, and cells with a higher percentage of reads mapping to the mitochondrial genome. For the $Mkl^{-/-}$ sample, cells were included if the number of unique genes detected per cell was between 500 and 3,500, and the

percentage of mitochondrial reads $<4\%$. For the $Casp8^{C362A/C362A} Mkl^{-/-}$ sample, cells were included if the number of unique genes detected per cell was between 300 and 3,000, and the percentage of mitochondrial reads $<4\%$. After filtering, we obtained 1,328 $Mkl^{-/-}$ cells and 874 $Casp8^{C362A/C362A} Mkl^{-/-}$ cells. The integration method from Seurat 3.0³⁷ was used to identify cell types common and unique to both samples. Dimensionality reduction was then performed on the normalized and scaled data, followed by clustering and visualization using Unified Manifold Approximation and Projection (UMAP). Cell types in different clusters were assessed using marker genes reported for different subsets in the adult mouse intestine³⁸.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Bulk and single-cell RNA-sequencing data are available in full through the GEO database (accession GSE132134). Source Data for Figs. 1–3 and Extended Data Figs. 1–4, 6–8 are provided with the paper. Other datasets generated during and/or analysed during the current study are available from the corresponding authors on reasonable request.

29. Mariathasan, S. et al. Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature* **430**, 213–218 (2004).
30. Mariathasan, S. et al. Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440**, 228–232 (2006).
31. Kayagaki, N. et al. Non-canonical inflammasome activation targets caspase-11. *Nature* **479**, 117–121 (2011).
32. Kayagaki, N. et al. Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling. *Nature* **526**, 666–671 (2015).
33. Newton, K. et al. Activity of protein kinase RIPK3 determines whether cells die by necroptosis or apoptosis. *Science* **343**, 1357–1360 (2014).
34. Murphy, J. M. et al. The pseudokinase MLKL mediates necroptosis via a molecular switch mechanism. *Immunity* **39**, 443–453 (2013).
35. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
36. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
37. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
38. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

Acknowledgements We thank M. Dempsey, T. Scholl, F. Gallardo and B. Torres for animal husbandry, J. Zhang, K.-H. Sun, S. Haller and members of the Genentech genetic analysis laboratory for technical assistance, and C. Print and W. Alexander for Vav^{cre} and $Mkl^{-/-}$ mice, respectively.

Author contributions K.N., K.E.W., A.M., D.L.D., M.S.S. and Z.M. designed and performed experiments, M.R.-G. generated $Casp8^{C362A/+}$ mice, Y.Z. and R.R. analysed RNA-sequencing data, J.D.W. analysed histological data and V.M.D. contributed to experimental design. K.N. wrote the paper with input from all authors.

Competing interests All authors are employees of Genentech.

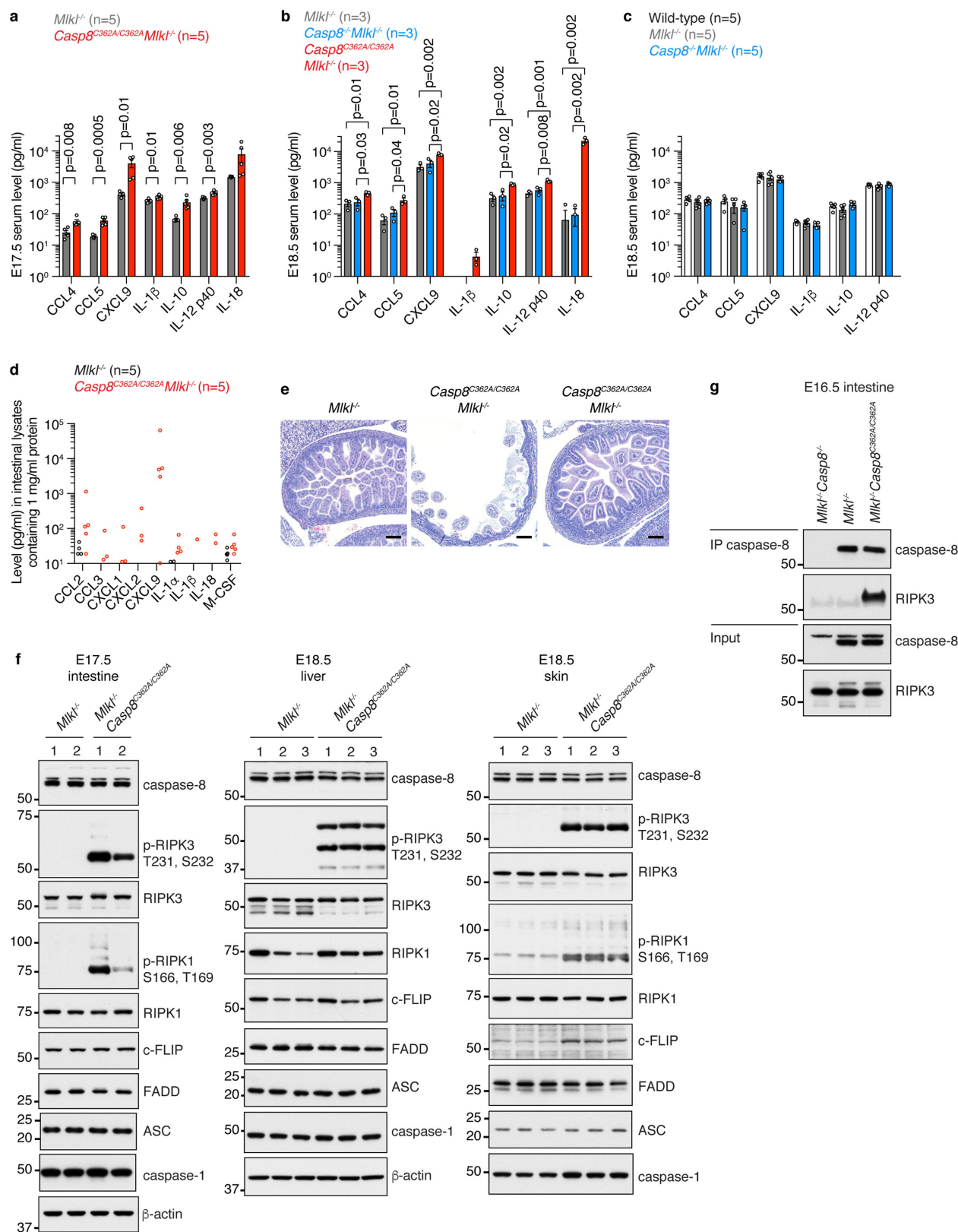
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1752-8>.

Correspondence and requests for materials should be addressed to K.N. or V.M.D.

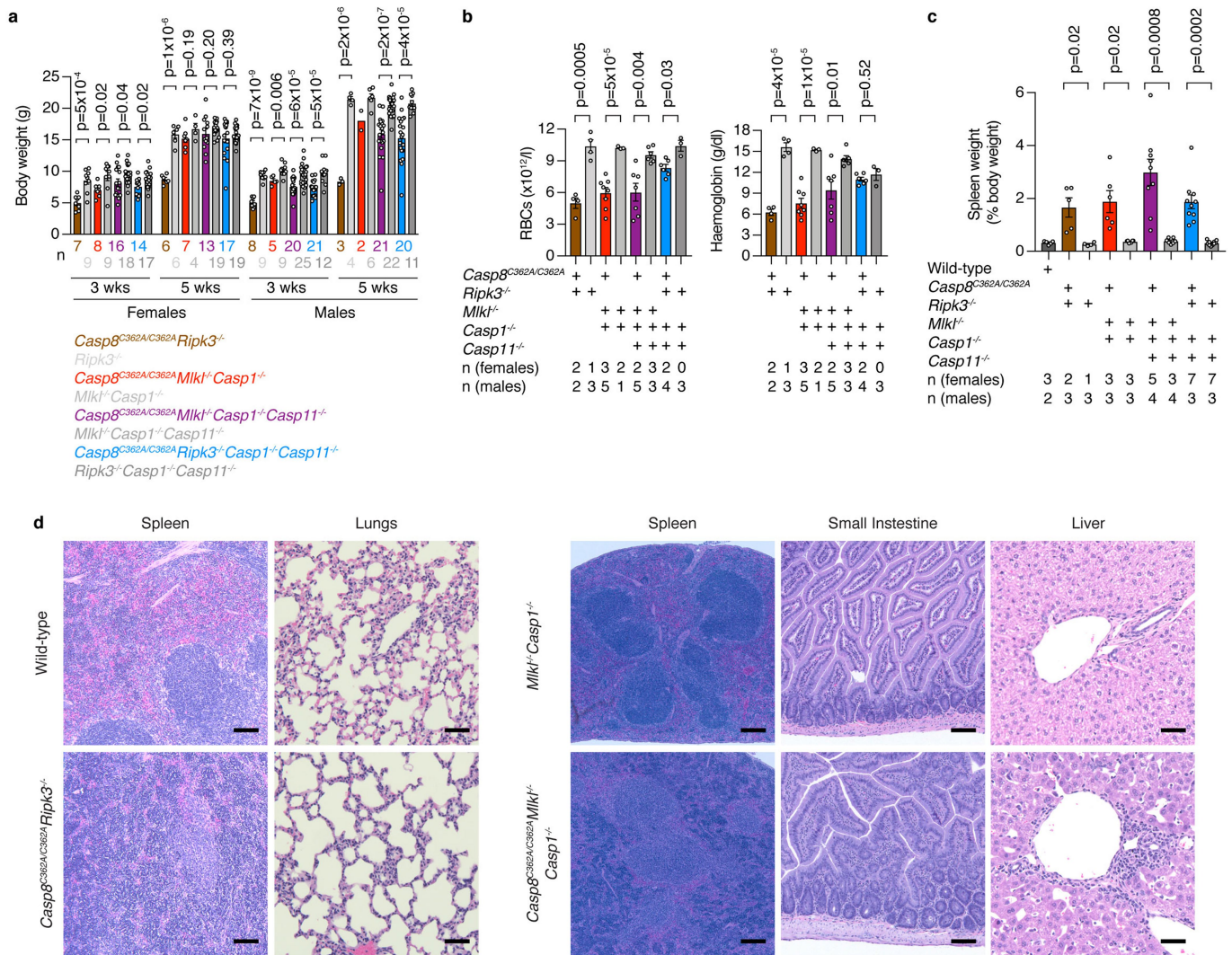
Peer review information Nature thanks Igor E. Brodsky, William Kaiser and Seamus J. Martin for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



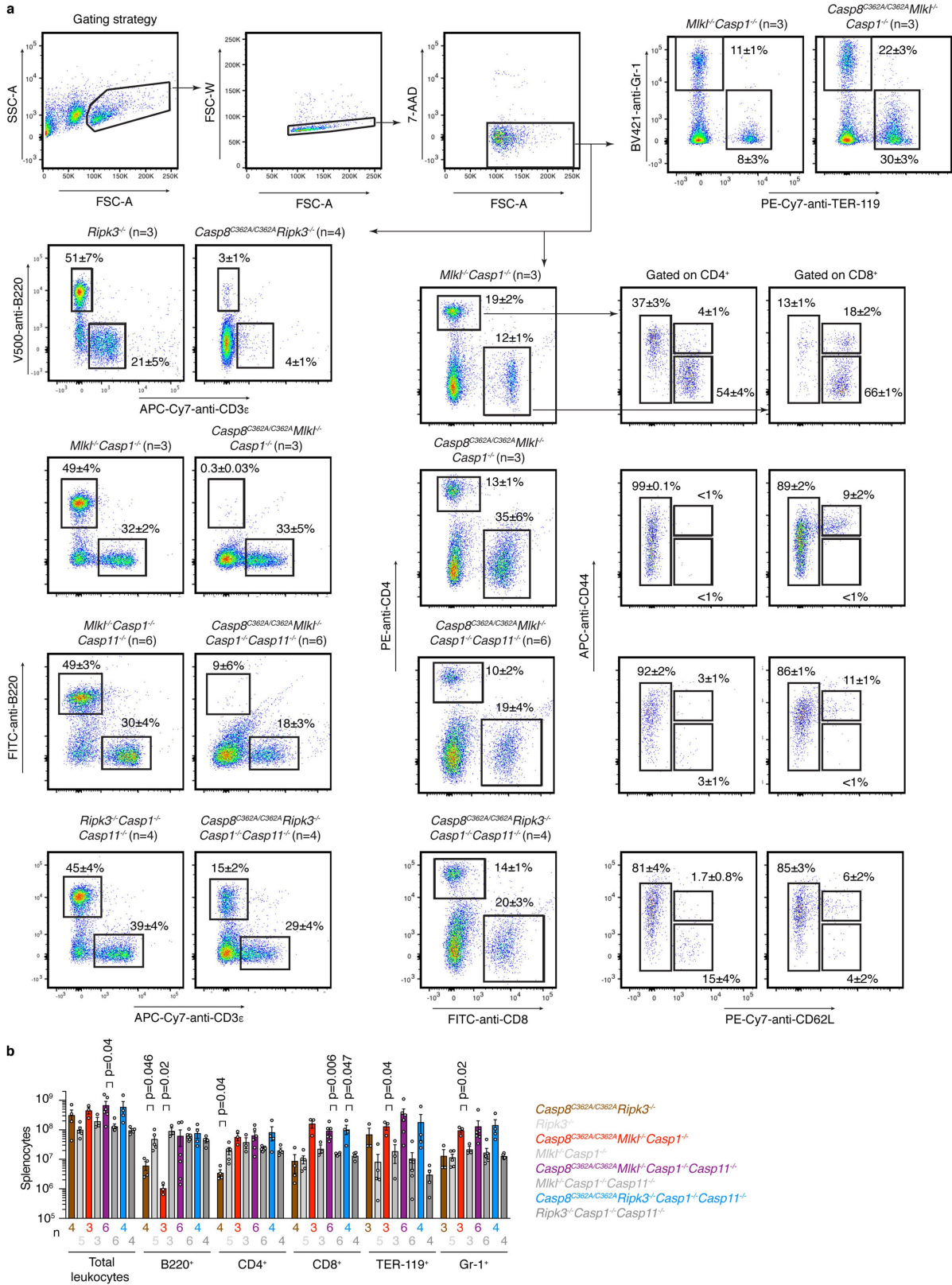
Extended Data Fig. 1 | Characterization of *Casp8^{C362A/C362A}Mik1^{-/-}* mice. **a–c**, Serum concentrations of cytokines and chemokines at E17.5 (**a**) or E18.5 (**b**, **c**). Circles, individual mice. Data are mean \pm s.e.m. *P* values are shown if *P* < 0.05; unpaired, two-tailed *t*-test. The concentration of IL-18 was below the limit of detection in **c**. **d**, Concentrations of cytokines and chemokines in the intestine at E17.5. Circles, individual mice. **e**, E17.5 intestines. Results representative of five *Mik1^{-/-}* mice (left) and five *Casp8^{C362A/C362A}Mik1^{-/-}* mice (middle and right), two of which

had intestinal atrophy (middle) and the other three showed a normal intestinal morphology. Scale bars, 100 μ m. **f**, Western blots of E17.5 intestine, E18.5 skin and E18.5 liver. Each lane is for a different mouse (*n* = 2 per genotype for intestine, 3 per genotype for skin and liver). **g**, Western blots of E16.5 intestine before (input) and after immunoprecipitation (IP) with anti-caspase-8 antibody. Results are representative of five independent experiments. For gel source data, see Supplementary Fig. 1.



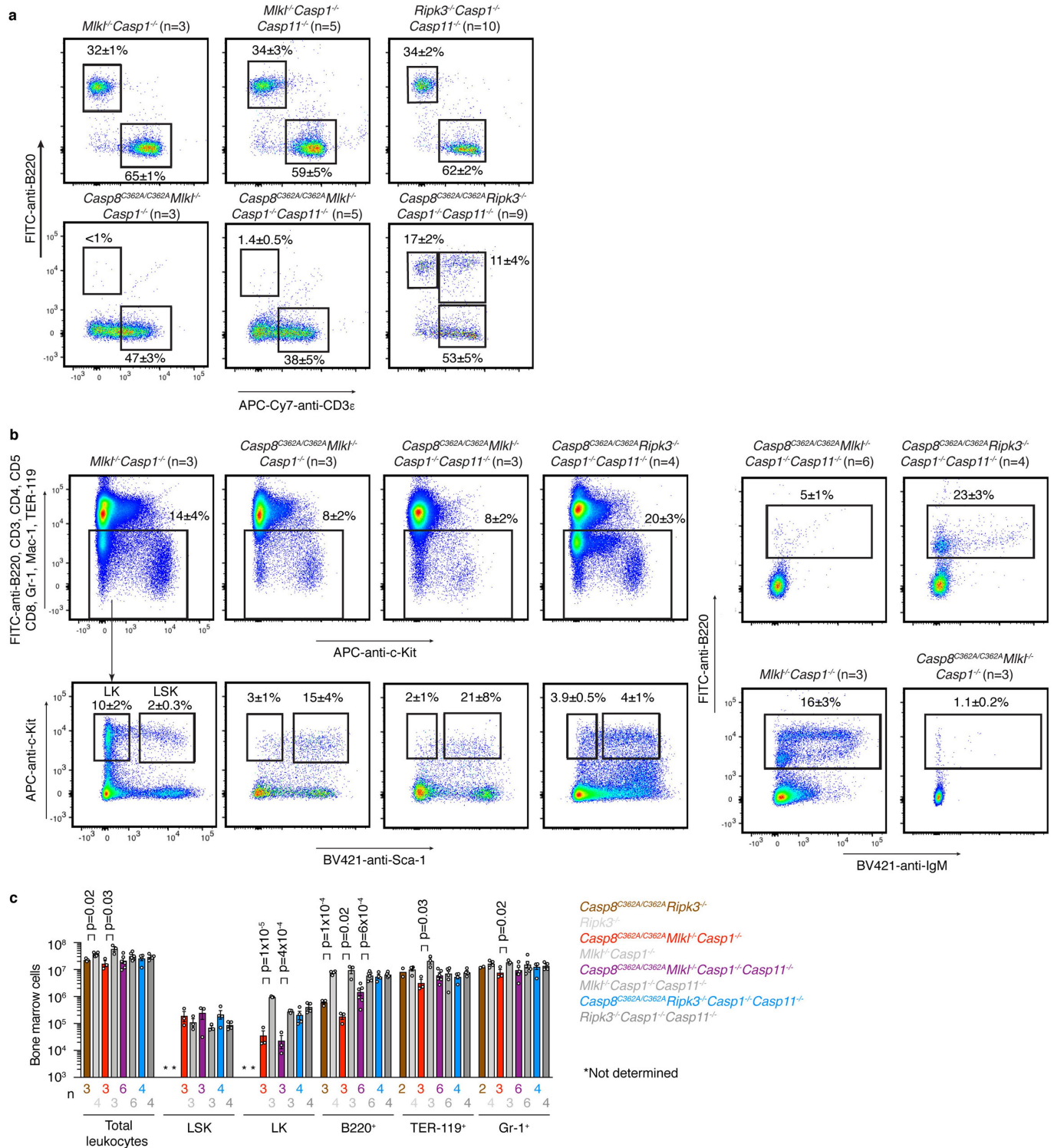
Extended Data Fig. 2 | Characterization of *Casp8^{C362A/C362A} Ripk3^{-/-}*, *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-}*, *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-} Casp11^{-/-}* and *Casp8^{C362A/C362A} Ripk3^{-/-} Casp1^{-/-} Casp11^{-/-}* mice. **a, Mouse body weights. Circles, individual mice. Data are mean \pm s.e.m. *P* values were calculated by unpaired, two-tailed *t*-test. **b**, Red blood cells (RBCs) and haemoglobin in mouse peripheral blood at 5–14 weeks. Circles, individual mice. Data are mean \pm s.e.m. *P* values were calculated by unpaired, two-tailed *t*-test with Welch's correction.**

c, Spleen weight as a percentage of body weight for mice aged 6–43 weeks. Circles, individual mice. Data are mean \pm s.e.m. *P* values by unpaired, two-tailed *t*-test with Welch's correction. **d**, Sections of spleen, liver, lungs and small intestine. Results are representative of two wild-type, four *Casp8^{C362A/C362A} Ripk3^{-/-}*, three *Mlkl^{-/-} Casp1^{-/-}* and three *Casp8^{C362A/C362A} Mlkl^{-/-} Casp1^{-/-}* mice aged 3–8 weeks. Scale bars, 200 μ m (spleen), 100 μ m (intestine) or 50 μ m (liver and lungs).

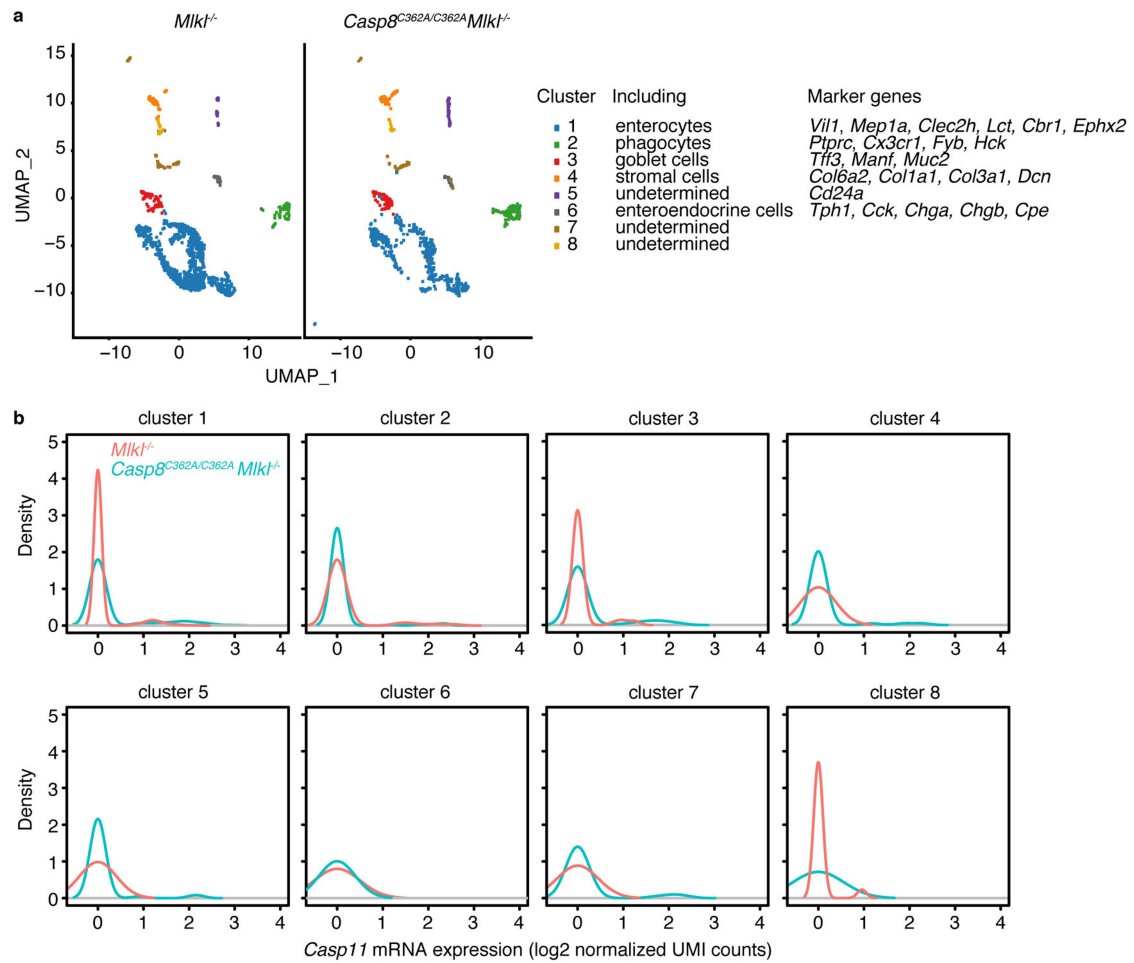


Extended Data Fig. 3 | Flow cytometry analysis of *Casp8^{C362A/C362A} Ripk3^{-/-}*, *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-}*, *Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-} Casp11^{-/-}* and *Casp8^{C362A/C362A} Ripk3^{-/-} Casp1^{-/-} Casp11^{-/-}* splenocytes. **a, Representative dot plots of splenocytes from mice aged 3–15 weeks (*Casp8^{C362A/C362A} Ripk3^{-/-}* and *Ripk3^{-/-}*) or 14–38 weeks (*Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-}*, *Mkl1^{-/-} Casp1^{-/-}*,**

Casp8^{C362A/C362A} Mkl1^{-/-} Casp1^{-/-} Casp11^{-/-}, *Mkl1^{-/-} Casp1^{-/-} Casp11^{-/-}*, *Casp8^{C362A/C362A} Ripk3^{-/-} Casp1^{-/-} Casp11^{-/-}* and *Ripk3^{-/-} Casp1^{-/-} Casp11^{-/-}*). Percentages are mean \pm s.e.m. **b**, Splenic leukocyte subsets. Circles, individual mice. Data are mean \pm s.e.m. *P* values are shown if *P* < 0.05; unpaired, two-sided *t*-test.

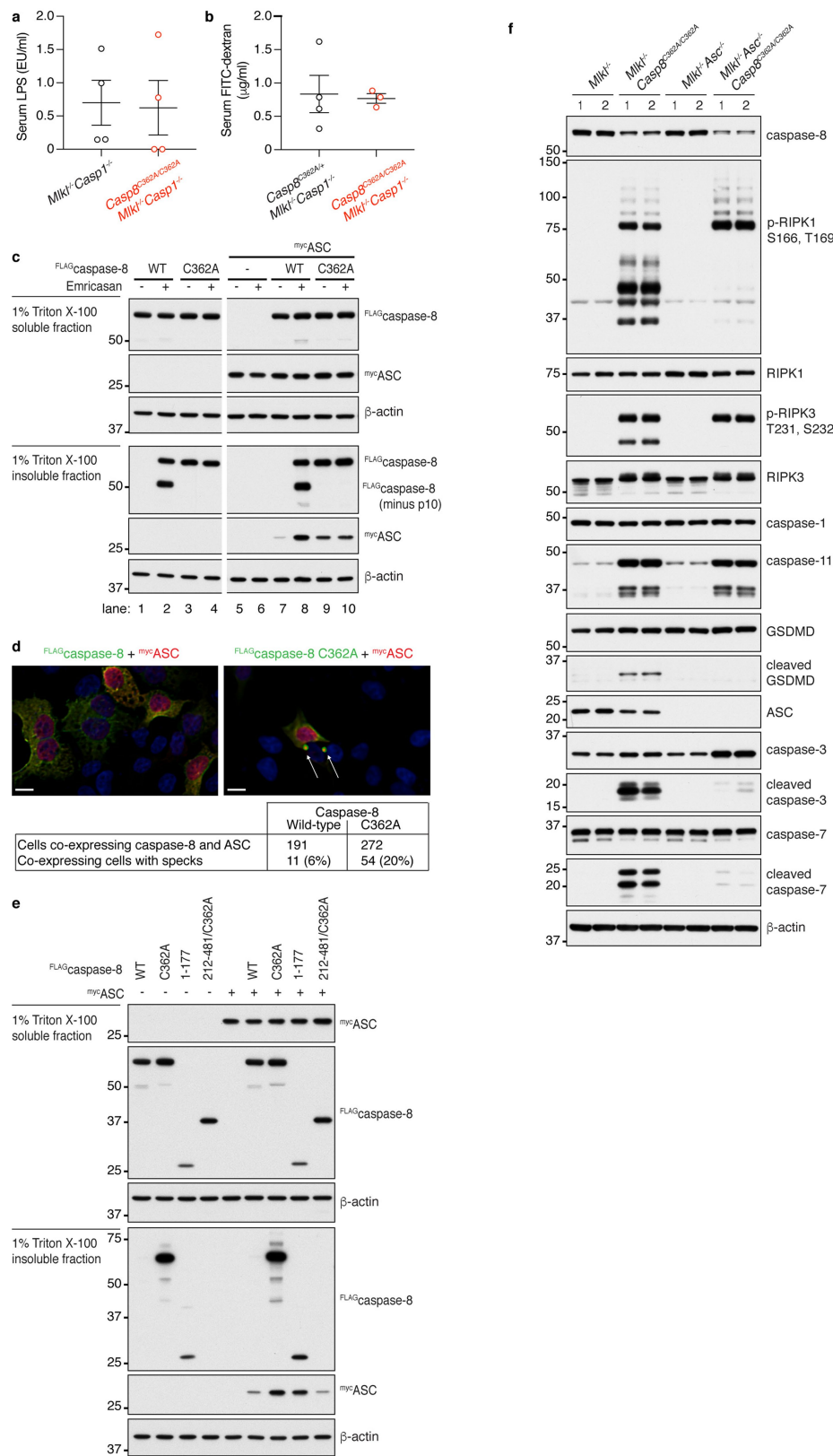


marrow leukocyte subsets. Circles, individual mice. Data are mean \pm s.e.m. P values are shown if $P < 0.05$; unpaired, two-sided t-test. Mice in **a–c** were aged 3–15 weeks (*Casp8*^{C362A/C362A} *Ripk3*^{-/-} and *Ripk3*^{-/-}) or 14–38 weeks (*Casp8*^{C362A/C362A} *Mlkt*^{-/-} *Casp1*^{-/-}, *Mlkt*^{-/-} *Casp1*^{-/-}, *Casp8*^{C362A/C362A} *Mlkt*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-}, *Mlkt*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-}, *Casp8*^{C362A/C362A} *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-} and *Ripk3*^{-/-} *Casp1*^{-/-} *Casp11*^{-/-}).



Extended Data Fig. 5 | Single-cell RNA sequencing of E18.5 intestines.
a, UMAP plots of E18.5 intestines analysed by single-cell RNA sequencing (*Mik1*^{-/-}, *n* = 1,328 cells; *Casp8*^{C362A/C362A} *Mik1*^{-/-}, *n* = 874 cells). **b**, Density plots

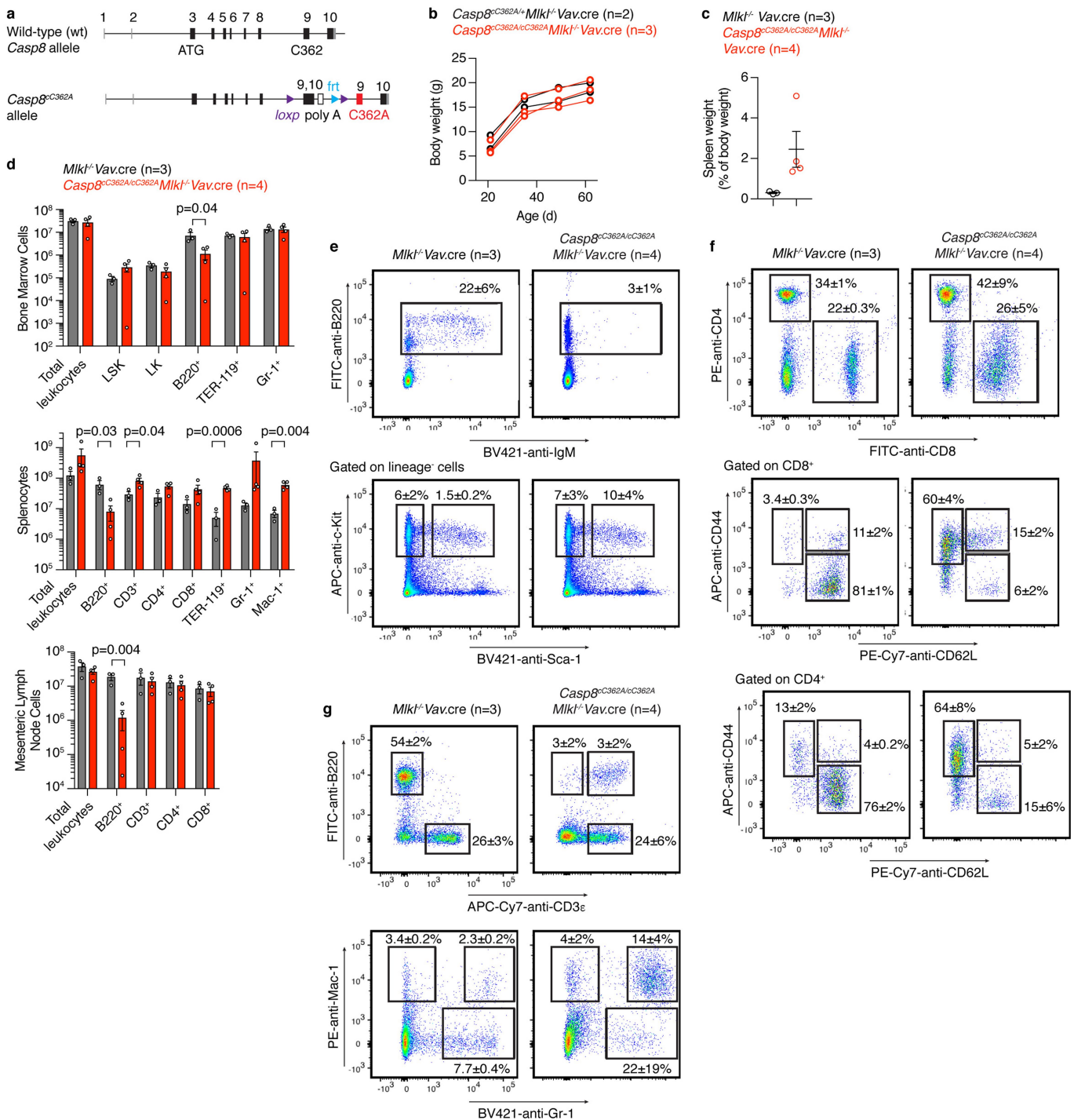
indicate expression of *Casp11* mRNA by cells in the eight clusters shown in **a**. Density corresponds to smoothed values for cell number. UMI, unique molecular identifier.



Extended Data Fig. 6 | See next page for caption.

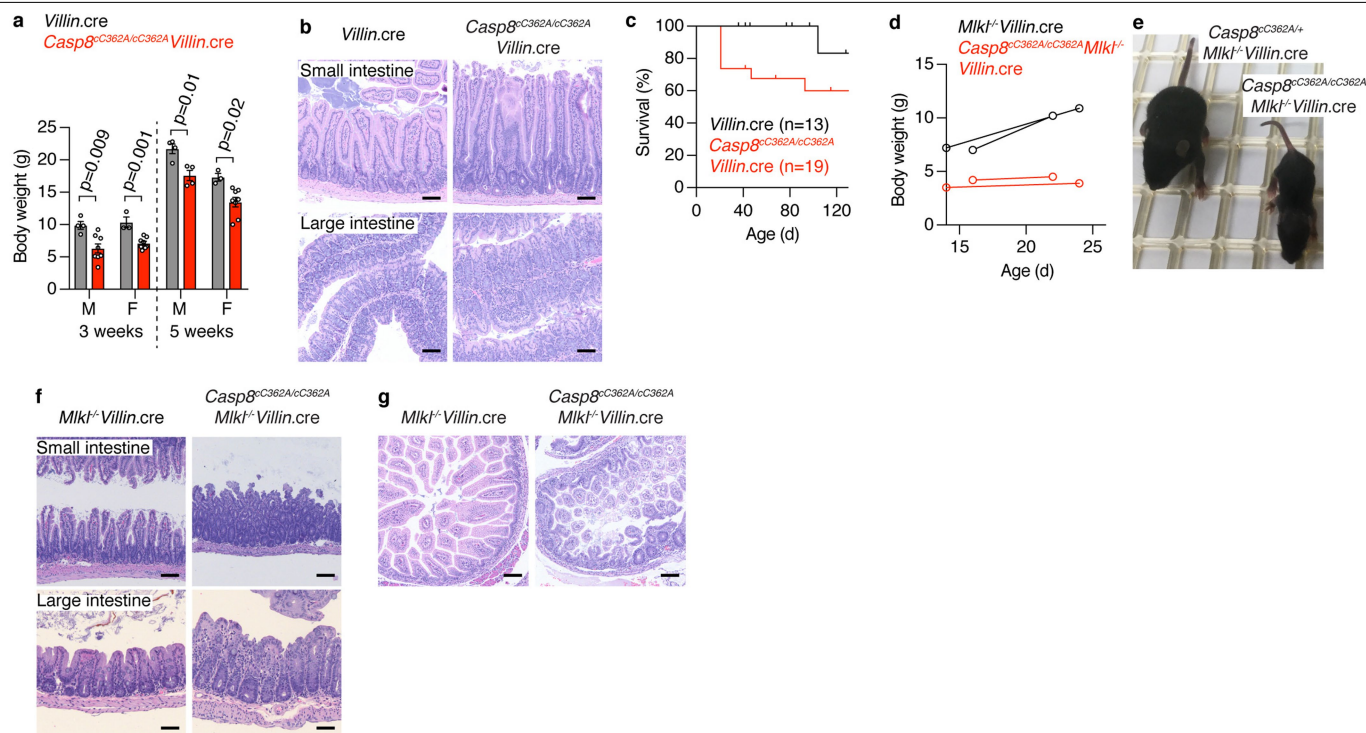
Extended Data Fig. 6 | The pro-domain of caspase-8 is necessary and sufficient for enhanced aggregation of ASC. **a**, Serum LPS for 7-day-old mice. Circles, individual mice ($n = 4$ per genotype). Data are mean \pm s.e.m. $P = 0.89$; unpaired, two-tailed t -test with Welch's correction. **b**, FITC-labelled dextran in the serum of 3–4-week-old mice at 5 h after gavage dosing with FITC-dextran. Circles, individual mice ($n = 3$, $Casp8^{C362A/C362A} Mkl^{-/-} Casp1^{-/-}$; $n = 4$, $Casp8^{C362A/+} Mkl^{-/-} Casp1^{-/-}$). Data are mean \pm s.e.m. $P = 0.83$; unpaired, two-tailed t -test with Welch's correction. **c**, Western blots of HEK293T cells transfected with caspase-8 and/or ASC. Results are representative of two independent experiments. Blotting of the loading control β -actin was performed after Flag. **d**, HEK293T cells labelled for transfected ASC (red), transfected caspase-8

(wild-type caspase-8 or mutant CASP8(C362A); green) and DNA (blue). The table indicates the proportion of cells co-expressing ASC and caspase-8 that contained speck-like structures (indicated by arrows in the image). Scale bars, 10 μ m. Results representative of two independent experiments. **e**, Western blots of HEK293T cells transfected with ASC and different caspase-8 mutants and then lysed with 1% Triton X-100. Results are representative of two independent experiments. Blotting of the loading control β -actin was performed after Flag (soluble fraction) or Myc (insoluble fraction). **f**, Western blots of E18.5 intestines. Lanes, different embryos ($n = 2$ per genotype). Blotting of the loading control β -actin was performed after phosphorylated (p-)RIPK1. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 7 | Characterization of mice expressing CASP8(C362A) only in haematopoietic and endothelial cells. a, Schematic showing the organization of the *Casp8^{cc362A}* conditional allele. **b**, Body weights of female littermates (*Casp8^{cc362A/+}Mik1^{-/-}Vav^{cre}*, n=2; *Casp8^{cc362A/cc362A}Mik1^{-/-}Vav^{cre}*, n=3). **c**, Spleen weight as a percentage of body weight for mice aged 4–6 months. Circles, individual mice (*Mik1^{-/-}Vav^{cre}*, n=3; *Casp8^{cc362A/cc362A}Mik1^{-/-}Vav^{cre}*, n=4). Lines, mean \pm s.e.m. *P*=0.09, unpaired, two-tailed *t*-test with Welch's

correction. **d**, Leukocyte numbers in mice aged 2–7 months. Bone marrow was from two femurs. Circles, individual mice (*Mik1^{-/-}Vav^{cre}*, n=2 males, 1 female; *Casp8^{cc362A/cc362A}Mik1^{-/-}Vav^{cre}*, n=1 male, 3 females). Data are mean \pm s.e.m. *P* values were calculated by unpaired, two-tailed *t*-test. **e–g**, Flow cytometry analysis of bone marrow (**e**), spleen (**f**) and mesenteric lymph node (**g**) cells from the mice in **d**. Samples analysed using the gating strategy shown in Extended Data Fig. 3.



Extended Data Fig. 8 | Characterization of mice expressing CASP8(C362A) only in intestinal epithelial cells. a, Mouse body weights. Circles, individual mice (*Villin^{cre}*, $n = 3$ females, 4 males; *Casp8^{C362A/cC362A}* *Mkt^{-/-}* *Villin^{cre}*, $n = 8$ females, 8 males at 3 weeks, or 8 females, 4 males at 5 weeks). Data are mean \pm s.e.m. P values by unpaired, two-tailed t -test. **b**, Small and large intestines of mice aged 19 weeks. Results representative of five *Casp8^{C362A/cC362A}* *Villin^{cre}* mice aged 3–19 weeks. Scale bars, 100 μ m. **c**, Kaplan–Meier survival curves of *Villin^{cre}* ($n = 13$) and *Casp8^{C362A/cC362A}* *Villin^{cre}* ($n = 19$) littermates.

$P = 0.06$, two-sided Gehan–Breslow–Wilcoxon test. **d**, Body weights of littermates ($n = 2$ per genotype). **e**, Littermate females aged 11 days. Results representative of two mice of each genotype. **f**, Small and large intestines of mice aged 4 weeks. Scale bars, 100 μ m (small intestine) or 50 μ m (large intestine). Results are representative of two mice of each genotype. **g**, Small intestine from 1-day-old littermates. Scale bars, 100 μ m. Similar histological changes were observed in two out of three newborn (P0) *Casp8^{C362A/cC362A}* *Mkt^{-/-}* *Villin^{cre}* pups.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Flow cytometry data acquired using BD FACSDiva software v8.0. Immunofluorescence data captured using Leica Application Suite v4.6.0. Fluorescence data captured using Softmax Pro v5.4.1. Luminex data acquired with Bio-Plex Manager 6.1.1. CBC panels acquired using Sysmex XT-2000iV, software version 00-13.
Data analysis	Flow cytometry data analysed using FlowJo 10.3. P-values calculated using Prism 7.0e. RNA sequencing data analysed using GSNAP and the voom/limma R package (both described in published literature - references provided in the manuscript). Single cell RNA sequencing data analysed with 10X Genomics Cell Ranger pipeline 3.0.2 and Seurat 3.0. Immunofluorescence data analysed with Huygens Professional v18.10 and Slidebook v6.0. Luminex data analysed with Bio-Plex Manager 6.1.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data are available in full through the GEO database, accession GSE132134. Other datasets generated during and/or analysed during the current study are available from the corresponding authors on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. Tissues and cells from at least 3 animals per genotype were analysed to be sure differences were reproducible. Variability between wild-type animals in the ex vivo assays used in this study tends to be low, so n=3 is the accepted norm in the field.
Data exclusions	No data were excluded from analyses.
Replication	Whenever possible, readouts were performed with at least 3 animals of a given genotype and all attempts at replication were successful.
Randomization	Groups were determined by genotype, not by any form of treatment, so randomization was not applicable.
Blinding	ASC speck images in fig. 3b were assessed by pathologist/author JDW blinded as to genotype. He successfully grouped slides according to genotype based on the IF images. Similarly, specks containing both caspase-8 and ASC in extended data fig. 6d were counted by author MSS blinded as to whether cells received wild-type caspase-8 or caspase-8(C362A). All other data were collected knowing the genotypes of the animals.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Antibodies used for flow cytometry were from BD Biosciences (2B8 anti-c-Kit, cat#553356, lot#7251822; IM7 anti-CD44, cat#559250, lot#30900; 145-2C11 anti-CD3, cat#557596, lot#7201885, or cat#553061, lot#41936; RB6-8C5 anti-Gr-1, cat#562709, lot#716628, or cat#553127, lot#3113790; R6-60.2 anti-IgM, cat#562595, lot#7277689; D7 anti-Sca-1, cat#562729, lot#7177519; RA3-6B2 anti-B220, cat#553088, lot#3352893 or cat#561226, lot#7048735; H129.19 anti-CD4, cat#553651, lot#40168, or cat#553653, lot#28488; 53-7.3 anti-CD5, cat#553021, lot#6294641; 53-6.7 anti-CD8, cat#553031, lot#3177592; M1/70 anti-Mac-1, cat#553310, lot#3010776, or cat#553311, lot#8018675; TER-119 anti-TER-119, cat#557915, lot#3253783; MEL-14 anti-CD62L, cat#560516, lot#7348721; 2.4G2 anti-CD16/CD32, cat#553142, lot#M063276). Conjugated antibodies were used at 5-10 ug/ml, whereas 2.4G2 anti-CD16/CD32 was used at 1 ug/ml.

Antibodies used for western blots were from Genentech (1G6 anti-mouse RIPK3 used at 0.25 ug/ml; GEN135-35-9 anti-mouse phospho-RIPK3 Thr231, Ser232 used at 1 ug/ml; 10C7 anti-mouse RIPK1 used at 0.5 ug/ml; GEN150-33-4 anti-mouse phospho-RIPK1 Ser166, Thr169 used at 1 ug/ml; 1.28E12 anti-mouse FADD used at 0.5 ug/ml; 2.21H2 anti-mouse c-FLIP used at 0.5 ug/ml; 4B4 anti-mouse caspase-1 used at 1 ug/ml; 8E4 anti-mouse ASC used at 0.125 ug/ml; GN20-13 anti-GSDMD used at 0.125 ug/ml), Enzo Life Sciences (1G12 anti-mouse caspase-8, cat#ALX-804-447-C100, lot#11131404 used at 0.2 ug/ml), EMD Millipore (anti-MLKL, cat#MABC604, lot#3083064 used at 0.5 ug/ml), BD Biosciences (38/RIP anti-RIPK1, cat#610459, lot#5177938 used at 0.25 ug/ml), Novus Biologicals (17D9 anti-caspase-11, cat#NB120-1045A, Lot#A-9 used at 0.5 ug/ml; anti-RIPK3, cat#NBP1-77299, lot#8337-1502 used at 0.25 ug/ml), MP Biomedicals (C4 anti-beta actin, cat#69100, lot#QR14180 diluted 1/20,000), Cell Signaling Technology (anti-RIPK1 cat#3493, lot#3, diluted 1/3,000; cleaved GSDMD, cat# 50928, lot#1 diluted 1/1,000; caspase-7, cat#9492, lot#6 diluted 1/3,000; cleaved caspase-7, cat#9491, lot#7 diluted 1/3,000; caspase-3, cat#9662, lot#18 diluted 1/3,000; cleaved caspase-3, cat#9664, lot#21 diluted 1/1,000), Sigma (M2 anti-Flag, cat#A8592, lot#SLBD9930 used at 0.05 ug/ml), Genetex (anti-myc polyclonal, cat#GTX21261, lot#821805083 used at 0.1 ug/ml), and Jackson

Immunoresearch (HRP-anti-mouse, cat#115-035-174; lot#139280; HRP-anti-rabbit, cat#211-032-171; lot#120918; HRP-anti-rat, cat#112-035-175; lot#115169; all diluted 1/10,000).

Mouse caspase-8 was immunoprecipitated with Abcam anti-caspase-8 antibody (cat#ab138485, lot#GR294765-3).

Antibodies used for immunofluorescence were from AdipoGen (anti-ASC polyclonal, cat#AG-25B-0006, lot#A28751704 diluted 1/500), Sigma (M2 anti-Flag, cat#F3165, lot#SLBT6752 diluted 1/5,000), Invitrogen (A488 anti-mouse, cat#A11029, lot 57465A diluted 1/500), and Jackson ImmunoResearch (Cy5-anti-rabbit, cat#711-176-152, lot 88587 diluted 1/500).

Validation

1G6 anti-mouse RIPK3 was validated for WB using wild-type and Ripk3^{-/-} mouse tissues in Newton et al (2004) Mol. Cell. Biol. 24:1464-1469.

GEN135-35-9 anti-mouse phospho-RIPK3 T231, S232 was validated for WB and IHC using wild-type (+/- phosphatase treatment) and Ripk3^{-/-} mouse samples in Newton et al (2016) Nature 540:129-133. Phosphospecificity also confirmed using ectopic mouse RIPK3, comparing wild-type and RIPK3(T231A,S232A).

The Novus Biologicals anti-RIPK3 antibody was validated for WB using wild-type and Ripk3 knockdown mouse cells in Lim et al (2019) Elife. 8:e44452. doi: 10.7554/eLife.44452.

10C7 anti-mouse RIPK1 and anti-RIPK1 (Cell Signaling Technology) were validated for WB using wild-type and Ripk1^{-/-} mouse cells.

38/RIP anti-RIPK1 was validated for WB using wild-type and Ripk1^{-/-} mouse cells in Newton et al (2016) Nature 540:129-133.

GEN150-33-4 anti-mouse phospho-RIPK1 S166, T169 was validated for WB using mouse cells in Heger et al (2018) Nature 559:120-124. In addition, it WBs wild-type mouse RIPK1 overexpressed in 293T cells, but not mouse RIPK1(S166A,T169A).

1.28E12 anti-mouse FADD was validated for WB using cells from wild-type and Fadd^{-/-} Mkl1^{-/-} mice.

2.21H2 anti-mouse c-FLIP was validated for WB using cells from wild-type and 3xFlag-c-FLIP knock-in mice.

8E4 anti-mouse ASC was validated for WB using wild-type and Asc^{-/-} mouse cells in Mariathasan et al (2004) Nature 430:213-218.

4B4 anti-mouse caspase-1 was validated for WB using wild-type and Casp1^{-/-} mouse cells in Kayagaki et al (2011) Nature 479:117-121.

17D9 anti-caspase-11 was validated for WB using wild-type and Casp11^{-/-} mouse cells in Kayagaki et al (2011) Nature 479:117-121.

The ASC polyclonal antibody (AdipoGen) was validated for IF using wild-type and Asc^{-/-} mouse cells.

1G12 anti-mouse casp8 was validated for WB in O'Reilly et al (2004) Cell Death Differ 11:724-736, and using Casp8^{-/-} Mkl1^{-/-} mouse cells in Extended data fig. 1g (this study).

The Abcam anti-casp8 polyclonal was validated for IP using Casp8^{-/-} Mkl1^{-/-} cells (Extended data fig. 1g, this study).

GN20-13 anti-GSDMD antibody was validated for WB using wild-type and Gsdmd^{-/-} mouse cells in Cerqueira et al (2018) PLoS Pathog 14:e1007519.

The EMD Millipore anti-MLKL antibody was validated for WB using Mkl1^{-/-} mouse tissues (Fig. 4, this study).

Validation data for commercial antibodies are available on vendor websites.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	293T cells (ATCC CRL-3216).
Authentication	Cells not authenticated.
Mycoplasma contamination	Cells negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	Not used.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	All mice (Mus musculus) were maintained on a C57BL/6N genetic background, with the exception of the villin.cre transgenic mice used for initial crosses, which were maintained on a C57BL/6J genetic background. Strains included Ripk1 ^{+/+} (Newton et al. 2014 Science 343:1357-1360), Ripk3 ^{-/-} (Newton et al. 2014 Science 343:1357-1360), Ripk3 RHIM/RHIM (Newton et al 2016 Nature 540:129-133), Casp8 ^{+/+} (Newton et al. 2014 Science 343:1357-1360), Casp8 ^{+/+} /C362A (Newton et al. 2019 Nature, in the press), Casp8 cC362A/cC362A (this study; extended data fig. 7a), Fadd ^{+/+} (Newton et al. 2019 Nature, in the press), Mkl1 ^{-/-} (Murphy et al. 2013 Immunity 39:443-453), Nlrp3 ^{-/-} (Mariathasan et al. 2006 Nature 440:228-232), Asc ^{-/-} (Mariathasan et al. 2004 Nature 430:218-218), Casp1 ^{-/-} (Kayagaki et al. 2015 Nature 526:666-671), Casp11 ^{-/-} (Kayagaki et al. 2011 Nature 479:117-121), Casp1 ^{-/-} /Casp11 ^{-/-} (Kayagaki et al. 2011 Nature 479:117-121), Villin.cre (Madison et al. 2002 J Biol Chem 277:33275-33283), and Vav.cre (Georgiades et al. 2002 Genesis 34:251-256). Mice of both sexes were analysed ranging in age from E16.5 through 8 months. Specifics are provided in each figure legend.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	All mouse studies were approved by the Genentech institutional animal care and use committee (IACUC).

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Bone marrow cells were flushed from two femurs after cutting off the ends of the bones. Splenocyte and lymph node single cell suspensions were prepared by gently pressing tissue through a 40 um cell strainer with the rubber plunger from a 3 ml syringe. An aliquot of cells was stained with acridine orange and DAPI (solution 18; Chemometec) and the viable cell concentration determined using a Nucleoview NC-250 cell counter (Chemometec). An equivalent number of viable cells in each sample were then stained for flow cytometry.
Instrument	BD FACSCantoll (BD Biosciences)
Software	Data was acquired using BD FACSDiva software v8.0, and analysed using FlowJo 10.3.
Cell population abundance	No sorting was performed.
Gating strategy	Leukocytes were identified based on their forward scatter (FSC-A) and side scatter (SSC-A) profiles. Dead cells that stained with 7-AAD (BD Biosciences), plus doublets, identified by their FSC-A versus FSC-W profiles, were excluded from analyses.
<input checked="" type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Caspase-8 is the molecular switch for apoptosis, necroptosis and pyroptosis

<https://doi.org/10.1038/s41586-019-1770-6>

Received: 18 February 2019

Accepted: 15 October 2019

Published online: 20 November 2019

Melanie Fritsch¹, Saskia D. Günther¹, Robin Schwarzer², Marie-Christine Albert¹, Fabian Schorn¹, J. Paul Werthenbach¹, Lars M. Schiffmann^{1,3}, Neil Stair^{1,2}, Hannah Stocks¹, Jens M. Seeger¹, Mohamed Lamkanfi^{4,5}, Martin Krönke¹, Manolis Pasparakis^{2,6} & Hamid Kashkar^{1,6*}

Caspase-8 is the initiator caspase of extrinsic apoptosis^{1,2} and inhibits necroptosis mediated by RIPK3 and MLKL. Accordingly, caspase-8 deficiency in mice causes embryonic lethality³, which can be rescued by deletion of either *Ripk3* or *Mlkl*^{4–6}. Here we show that the expression of enzymatically inactive CASP8(C362S) causes embryonic lethality in mice by inducing necroptosis and pyroptosis. Similar to *Casp8*^{−/−} mice^{3,7}, *Casp8*^{C362S/C362S} mouse embryos died after endothelial cell necroptosis leading to cardiovascular defects. MLKL deficiency rescued the cardiovascular phenotype but unexpectedly caused perinatal lethality in *Casp8*^{C362S/C362S} mice, indicating that CASP8(C362S) causes necroptosis-independent death at later stages of embryonic development. Specific loss of the catalytic activity of caspase-8 in intestinal epithelial cells induced intestinal inflammation similar to intestinal epithelial cell-specific *Casp8* knockout mice⁸. Inhibition of necroptosis by additional deletion of *Mlkl* severely aggravated intestinal inflammation and caused premature lethality in *Mlkl* knockout mice with specific loss of caspase-8 catalytic activity in intestinal epithelial cells. Expression of CASP8(C362S) triggered the formation of ASC specks, activation of caspase-1 and secretion of IL-1β. Both embryonic lethality and premature death were completely rescued in *Casp8*^{C362S/C362S} *Mlkl*^{−/−} *Asc*^{−/−} or *Casp8*^{C362S/C362S} *Mlkl*^{−/−} *Casp1*^{−/−} mice, indicating that the activation of the inflammasome promotes CASP8(C362S)-mediated tissue pathology when necroptosis is blocked. Therefore, caspase-8 represents the molecular switch that controls apoptosis, necroptosis and pyroptosis, and prevents tissue damage during embryonic development and adulthood.

In addition to its role in apoptosis and necroptosis, recent in vitro studies have indicated that caspase-8 induces the production of cytokines by acting as a scaffolding protein and that this role is independent of its enzymatic activity^{9,10}. The scaffold function of caspase-8 was also shown to be involved in the double-stranded RNA (dsRNA)-induced activation of the NLRP3 inflammasome in macrophages¹¹. Additional studies indicate that the enzymatic activity of caspase-8 is required for the activation of NF-κB and secretion of cytokines in response to activated antigen receptors, Fc receptors or Toll-like receptors (TLRs), independently of cell death^{12,13}. To investigate the physiological role of the enzymatic activity of caspase-8, we generated knock-in mice that expressed catalytically inactive caspase-8 by mutating Cys362 in the substrate binding pocket to serine (C362S) (Extended Data Fig. 1a). Although heterozygous *Casp8*^{C362S/WT} mice were viable (Extended Data Fig. 1b), *Casp8*^{C362S/C362S} embryos died around embryonic day 11.5 (E11.5). Hyperaemia in the abdominal areas was detected in *Casp8*^{C362S/C362S} embryos (Fig. 1a), presumably owing to defects in vascular development, which

resembles the phenotype of *Casp8*^{−/−} embryos⁷. In order to address the role of caspase-8 in vascular development, we used *Tie2*^{cre} (also known as *Tek*^{cre}) mice, in which efficient Cre-mediated recombination is induced in all endothelial cells and most haematopoietic cells¹⁴. Loss of the catalytic activity of caspase-8 in *Casp8*^{C362S/fl} *Tie2*^{cre} mice or specific knockout of caspase-8 in the endothelial cells of *Casp8*^{fl/fl} *Tie2*^{cre} mice caused embryonic lethality at the same developmental stage as *Casp8*^{C362S/C362S} embryos (Extended Data Fig. 1c, d). *Casp8*^{C362S/fl} *Tie2*^{cre} and *Casp8*^{fl/fl} *Tie2*^{cre} embryos showed the same gross pathology associated with a decrease in yolk-sac vascularization (Fig. 1b and Extended Data Fig. 1d).

Specific loss of the catalytic activity of caspase-8 in epidermal keratinocytes or intestinal epithelial cells was achieved by crossing *Casp8*^{C362S/fl} mice with *Krt14*^{cre} (also known as *K14*^{cre}) mice¹⁵ or *Vill*^{cre} (also known as *Villin*^{cre}) mice¹⁶, respectively. Loss of the catalytic activity of caspase-8 in these two cell types caused similar pathologies to caspase-8 deficiency in these tissues. *Casp8*^{C362S/fl} *K14*^{cre} and *Casp8*^{fl/fl}

¹Institute for Medical Microbiology, Immunology and Hygiene (IMMIH), CECADE Research Center, University of Cologne, Cologne, Germany. ²Institute for Genetics, CECADE Research Center, University of Cologne, Cologne, Germany. ³Department of General, Visceral and Cancer Surgery, University of Cologne, Cologne, Germany. ⁴Department of Internal Medicine and Paediatrics, Ghent University, Ghent, Belgium. ⁵VIB Center for Inflammation Research, Ghent, Belgium. ⁶Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany.

*e-mail: h.kashkar@uni-koeln.de

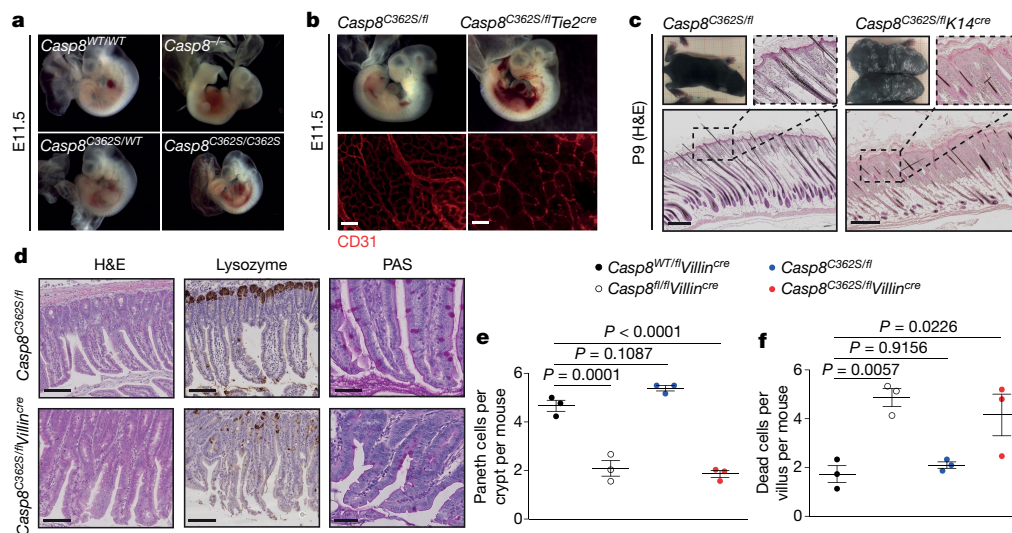


Fig. 1 | The enzymatic activity of caspase-8 is required to inhibit necroptosis.

a, b, Representative images of *Casp8*^{WT/WT} (*n* = 5), *Casp8*^{C362S/WT} (*n* = 8), *Casp8*^{C362S/C362S} (*n* = 3), *Casp8*^{-/-} (*n* = 3), *Casp8*^{C362S/WT} (*n* = 5) and *Casp8*^{C362S/WT}Tie2^{cre} (*n* = 5) mouse embryos at E11.5 (**a, b**, top). CD31 staining as endothelial marker of whole-mount yolk sacs (**b**, bottom). Scale bars, 100 μ m. **c,** Representative images of 9-day-old (P9) *Casp8*^{C362S/WT} (*n* = 3) and *Casp8*^{C362S/WT}K14^{cre} (*n* = 3) mice (top left) and skin sections stained with haematoxylin and eosin (H&E) (top

right, bottom). Scale bars, 100 μ m (magnification, top) and 300 μ m (bottom).

d, Ileal sections from 10-week-old *Casp8*^{C362S/WT} (*n* = 3) and *Casp8*^{C362S/WT}Villin^{cre} (*n* = 4) mice stained with H&E (left), for lysozyme (Paneth cells, middle) and periodic acid–Schiff (PAS) (right). Scale bars, 100 μ m. **e, f,** Count of Paneth cells (**e**) and dead IECs (**f**) (per crypt per mouse, *n* = 3). Dots, individual mice. Data are mean \pm s.e.m. One-way analysis of variance (ANOVA) followed by Dunnett's post-analysis.

K14^{cre} mice developed inflammatory skin lesions with focal epidermal thickening and scaling that appeared 5–7 days after birth and these lesions gradually increased in size and distribution and covered large cutaneous areas (Fig. 1c and Extended Data Fig. 1e). Histological skin analyses revealed epidermal hyperplasia and immune-cell infiltration into the dermis of *Casp8*^{C362S/WT}K14^{cre} and *Casp8*^{fl/fl}K14^{cre} mice as has previously been shown by transgenic overexpression of CASP8(C362S) in

epidermal keratinocytes¹⁷. Mice with intestinal epithelial cell (IEC)-specific loss of caspase-8 catalytic activity (*Casp8*^{C362S/WT}Villin^{cre}) developed ileitis at 8–10 weeks of age, associated with increased numbers of dying IECs, an altered distribution of goblet cells and loss of Paneth cells; a phenotype that is similar to that found in *Casp8*^{fl/fl}Villin^{cre} mice⁸ (Fig. 1d–f and Extended Data Fig. 2a, b). Together, these results demonstrate that the loss of the catalytic activity of caspase-8 causes similar

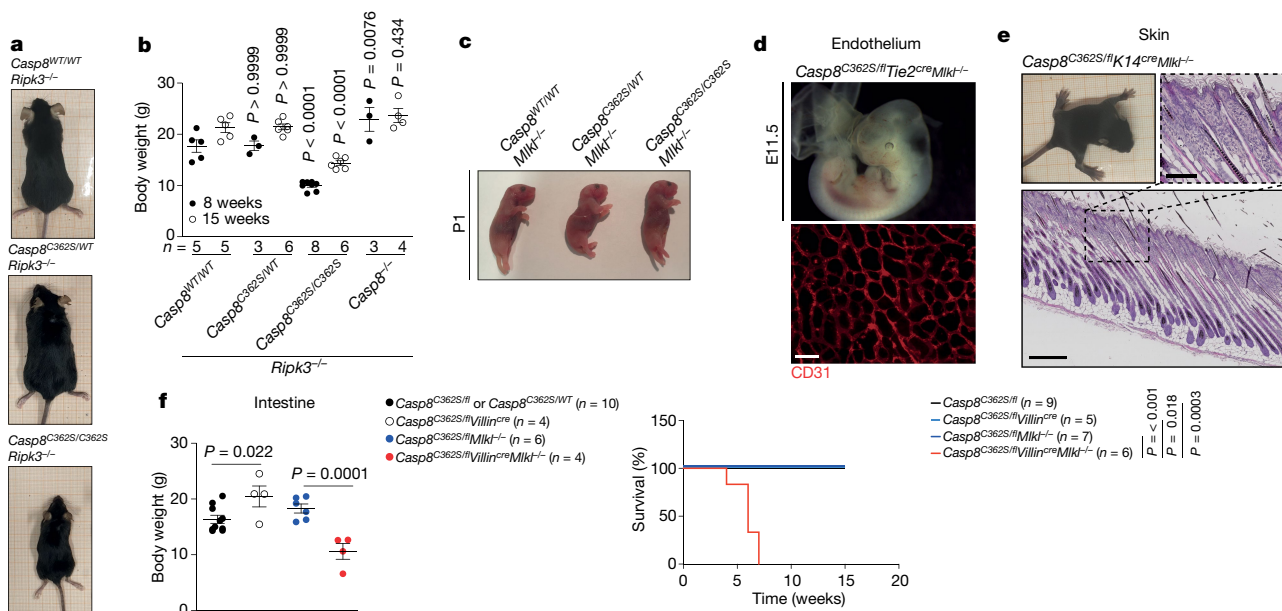


Fig. 2 | CASP8(C362S) induces necroptosis-independent tissue destruction.

a, Representative images of 8-week-old mice. **b,** Body weight of mice of the indicated ages. Dots, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis compared to the corresponding *Casp8*^{WT/WT} values. **c,** Representative images of 1-day-old *Mlkt*^{-/-} (*n* = 2), *Casp8*^{C362S/WT}*Mlkt*^{-/-} (*n* = 4) and *Casp8*^{C362S/C362S}*Mlkt*^{-/-} (*n* = 3) mouse neonates. **d,** Representative images of an embryo (*n* = 3) at E11.5 (top) and CD31 staining of a whole-mount

yolk sac (bottom). Scale bar, 100 μ m. **e,** Representative images of 9-day-old mice (*n* = 4) (top left) and skin sections stained with H&E (top right, bottom). Scale bars, 100 μ m (magnification, top) and 300 μ m (bottom). **f,** Body weight of 5-week-old mice (top). Dots, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis. Kaplan–Meier survival curves of mice as indicated (bottom). *P* values by two-sided log-rank test.

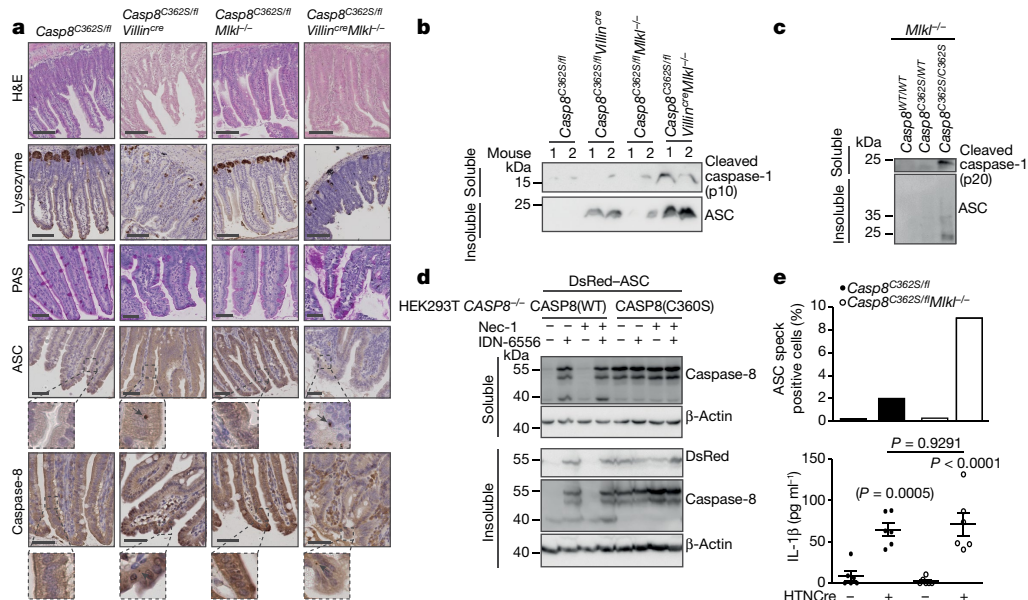


Fig. 3 | CASP8(C362S) activates the ASC inflammasome in IECs. **a**, Ileal sections of *Casp8*^{C362S/f1} (*n* = 4), *Casp8*^{C362S/f1}*Villin*^{cre} (*n* = 5), *Casp8*^{C362S/f1}*Mlkl*^{-/-} (*n* = 4) and *Casp8*^{C362S/f1}*Villin*^{cre}*Mlkl*^{-/-} (*n* = 5) 5-week-old mice with H&E, lysozyme, PAS, ASC and caspase-8 staining. Scale bars, 100 μ m (H&E and lysozyme) and 50 μ m (PAS, ASC and caspase-8). Arrows, ASC or caspase-8 aggregates. **b**, **c**, Western blot analysis of ileal lysates from two representative mice (**b**; mice are shown in **a**) or from one representative P1 mouse neonate (**c**; mice are shown in Fig. 2c) detecting cleaved caspase-1 and ASC. Lanes, individual mice. **d**, Western blot analysis of *CASP8*^{-/-} HEK293T cells (clone 1, Extended Data

Fig. 5e) transfected with DsRed-tagged human ASC (DsRed-ASC) either with human wild-type caspase-8 or *CASP8*(C360S) and treated with Nec-1 and/or IDN-6556 for 14 h, as indicated. Results representative of two individual experiments. **e**, ASC-speck-positive BMDMs (*n* = 100, one representative experiment) (top) and IL-1 β measurement in supernatants of BMDMs (bottom) after 24 h HTNCre treatment in biologically independent replicates (*n* = 6, representative of two individual experiments). Dots represent an individual biological replicates. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis to the corresponding values without HTNCre.

pathologies to those found in *Casp8*^{-/-} mice during embryonic development and adult tissue homeostasis.

To further characterize the cells that express *CASP8*(C362S), we isolated endothelial cells from *Casp8*^{C362S/f1} and *Casp8*^{WT/f1} mice and induced *Casp8* gene deletion in vitro using a cell-permeable active Cre protein (HTNCre)¹⁸ (Extended Data Fig. 3a). Endothelial cells that expressed *CASP8*(C362S) were unable to activate caspase-3 in response to TNF, but were sensitized to TNF-induced necroptosis, as the cytotoxic effect of TNF was associated with MLKL phosphorylation and was abolished by co-treatment with necrostatin-1 (Nec-1) (Extended Data Fig. 3b, c). Loss of RIPK3 or MLKL was previously shown to inhibit necroptosis and to prevent the embryonic lethality caused by caspase-8 knockout⁴⁻⁶. Similar to *Casp8*^{-/-}*Ripk3*^{-/-} mice, *Casp8*^{C362S/C362S}*Ripk3*^{-/-} mice survived weaning but showed markedly stunted growth and suffered from anaemia with distinct haematological abnormalities that led to splenomegaly (Fig. 2a, b and Extended Data Fig. 3d–g). MLKL deficiency¹⁹ did not rescue the lethality caused by the *CASP8*(C362S) mutation, leading to perinatal death of *Casp8*^{C362S/C362S}*Mlkl*^{-/-} mice (Fig. 2c and Extended Data Fig. 3h). Although *Casp8*^{C362S/C362S}*Mlkl*^{-/-} embryos were present at E13.5 in expected numbers and without any gross phenotypic alterations (Extended Data Fig. 4a), deletion of MLKL did not lead to *Casp8*^{C362S/C362S} animals reaching weaning age. Together, necroptosis deficiency did not fully lead to *Casp8*^{C362S/C362S} animals reaching adulthood, which suggests that the loss of the enzymatic activity of caspase-8 compromises perinatal development by additional, necroptosis-independent functions.

To dissect the necroptosis-independent role of *CASP8*(C362S) in different tissues, we assessed how MLKL deficiency affects the phenotypes that develop after specific expression of *CASP8*(C362S) in the endothelium (and haematopoietic cells), the skin or the intestinal epithelium. *Casp8*^{C362S/f1}*Tie2*^{cre}*Mlkl*^{-/-} mice survived weaning without gross phenotypic alterations yet developed splenomegaly at 8 weeks of age, similar to *Casp8*^{-/-}*Mlkl*^{-/-} mice⁶ (Fig. 2d and Extended Data Fig. 4b). These observations indicate that the enzymatic activity of caspase-8 is

required to inhibit necroptosis at early stages of embryonic development (E9–E12) that involve the formation of the cardiovascular system and placenta. Inflammatory skin lesions induced by *CASP8*(C362S) expression were not detected in MLKL-deficient *Casp8*^{C362S/f1}*K14*^{cre}*Mlkl*^{-/-} mice, indicating that the prevention of necroptosis by the intact enzymatic activity of caspase-8 is crucial for skin homeostasis (Fig. 2e and Extended Data Fig. 4c).

By contrast, *Casp8*^{C362S/f1}*Villin*^{cre}*Mlkl*^{-/-} mice showed reduced body weight and died at 4–8 weeks of age (Fig. 2f and Extended Data Fig. 4d). Histological analyses revealed intestinal inflammation in *Casp8*^{C362S/f1}*Villin*^{cre}*Mlkl*^{-/-} mice (at 5–7 weeks of age) that was characterized by pronounced villus atrophy, reduced numbers of Paneth cells, an altered distribution of goblet cells, the elongation of the crypts and hyperplasia in the ileum (Fig. 3a and Extended Data Fig. 5a). Therefore, MLKL deficiency not only could not rescue the ileitis that developed in *Casp8*^{C362S/f1}*Villin*^{cre} mice, but also strongly exacerbated the phenotype that caused premature death in these mice. These findings reveal that catalytically inactive caspase-8 triggers intestinal pathology in a necroptosis-independent way, which led us to search for caspase-8 catalytic activity-dependent mechanisms that prevent intestinal inflammation and that are distinct from its role in inhibiting necroptosis.

To further characterize the mechanisms that underlie the necroptosis-independent intestinal pathology found in *Casp8*^{C362S/f1}*Villin*^{cre}*Mlkl*^{-/-} mice, we examined the expression of a panel of secreted and soluble cytokines in ileal protein extracts (Extended Data Fig. 5b). This analysis revealed the pronounced secretion of IL-1 β and TNF in the ileum of *Casp8*^{C362S/f1}*Villin*^{cre}*Mlkl*^{-/-} mice. In particular, increased levels of IL-1 β were associated with proteolytic activation of the inflammatory caspase, caspase-1, which was detected in the soluble fraction of ileal lysates (Fig. 3b and Extended Data Fig. 5c). Caspase-1 is activated by canonical inflammasomes that involve members of the Nod-like receptor protein family such as NLRP3 and the inflammasome adaptor ASC that aggregates to form macromolecular ASC specks and serves

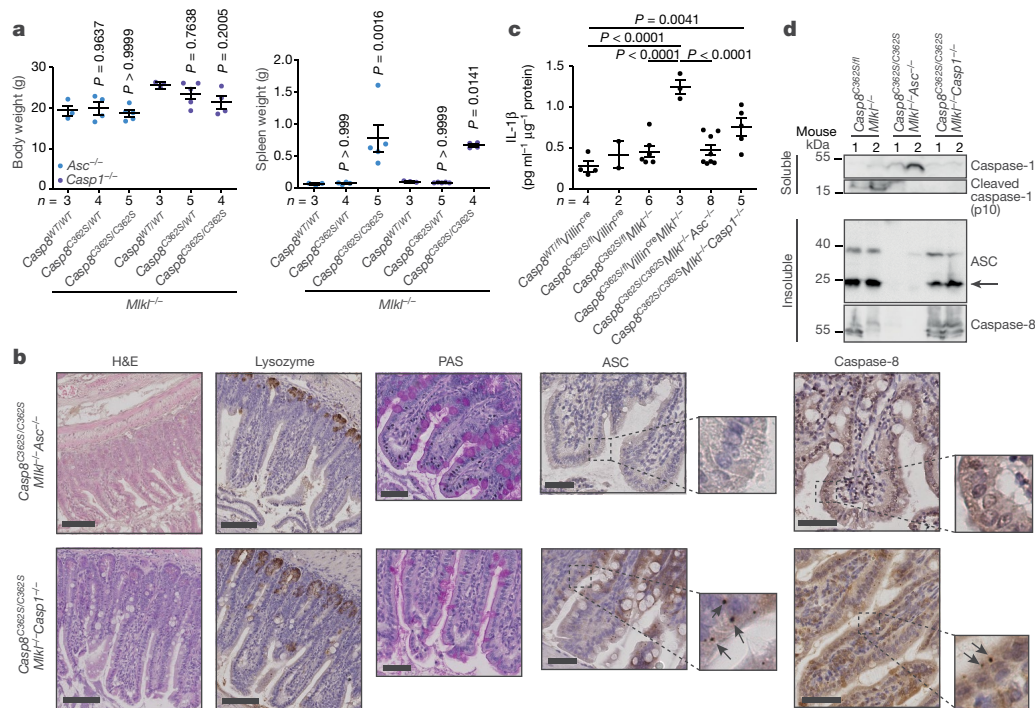


Fig. 4 | ASC or caspase-1 deficiency rescues embryonic lethality of mice expressing CASP8(C362S). **a**, Body and spleen weight of 8-week-old mice. Dots, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis compared to the corresponding *Casp8*^{WT/WT} values. **b**, Representative ileal sections from 5-week-old *Casp8*^{C362S/C362S}*Mlkl*^{-/-}*Asc*^{-/-} (*n* = 4) and *Casp8*^{C362S/C362S}*Mlkl*^{-/-}*Casp1*^{-/-} (*n* = 3) mice with H&E, lysozyme, PAS,

ASC and caspase-8 staining. Scale bars, 100 μ m (H&E and lysozyme) and 50 μ m (PAS, ASC and caspase-8). Arrows, ASC or caspase-8 aggregates. **c**, IL-1 β enzyme-linked immunosorbent assay (ELISA) in ileal lysates. Dots, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Turkey's post-analysis. **d**, Western blot analysis of ileal lysates from two representative mice (shown in b) detecting cleaved caspase-1, ASC and caspase-8. Lanes, individual mice.

as activation platform for caspase-1²⁰. Expression of CASP8(C362S) resulted in the aggregation of ASC in the insoluble fraction of ileal lysates from *Casp8*^{C362S/WT}*Villin*^{cre} mice and the aggregation was further exacerbated upon inhibition of necroptosis in *Casp8*^{C362S/WT}*Villin*^{cre}*Mlkl*^{-/-} mice (Fig. 3b). Furthermore, IL-1 β secretion, caspase-1 activation and ASC aggregation was also detected in intestinal lysates that were derived from *Casp8*^{C362S/C362S}*Mlkl*^{-/-} neonates (Fig. 3c and Extended Data Fig. 5b, d). Immunostaining revealed the formation of ASC specks in IECs that express CASP8(C362S), in particular, when necroptosis was blocked in *Casp8*^{C362S/WT}*Villin*^{cre}*Mlkl*^{-/-} mice (Fig. 3a). These results suggest that expression of catalytically inactive caspase-8 causes activation of the inflammasome both in neonate and adult mouse intestines.

To mechanistically characterize the ability of catalytically inactive caspase-8 to induce the formation of ASC specks, we established two independent human cell lines that lacked caspase-8 expression (Extended Data Fig. 5e) and ectopically expressed human caspase-8 and human ASC after transient transfection. Only upon ectopic expression of enzymatically inactive human CASP8(C360S), but not wild-type caspase-8, ASC aggregates were increasingly detected in the Triton X-100 insoluble cellular fractions (Fig. 3d and Extended Data Fig. 5f). Wild-type caspase-8 only induced ASC aggregation when transfected cells were additionally treated with the pan-caspase inhibitor IDN-6556 (emricasan). Immunofluorescence staining of caspase-8 in transfected cells revealed that most of the CASP8(C360S) and wild-type caspase-8 protein co-localized with cytoplasmic ASC aggregates in IDN-6556-treated cells (Extended Data Fig. 6a). Notably, CASP8(C360S) was detected in the soluble and insoluble cellular fractions independently of IDN-6556 treatment, whereas full-length wild-type caspase-8 was expressed at low levels in untreated cells, the expression of which markedly increased in the presence of the IDN-6556 (Fig. 3d and Extended Data Fig. 4f). Immunoprecipitation of the overexpressed caspase-8 variants in total cell lysates from *CASP8*^{-/-} HEK293T cells indicated that

the human caspase-8 mutant CASP8(C360S) interacts with human ASC (Extended Data Fig. 6b). Notably, overexpression of wild-type caspase-8 in cells resulted in its autocleavage and a reduction in the protein level of caspase-8. Only when HEK293T cells were treated with IDN-6556, comparable amounts of caspase-8 could be detected, the expression of which, in turn, induced ASC aggregation and co-immunoprecipitation. Thus, overexpression studies suggest that the expression of inactive caspase-8 alone is sufficient for the formation of ASC specks.

In addition to transfection studies, isolated bone-marrow-derived macrophages (BMDMs) from *Casp8*^{C362S/WT} and *Casp8*^{C362S/WT}*Mlkl*^{-/-} mice were exposed to HTNcre to induce the deletion of the *Casp8* floxed alleles in vitro. The deletion of the floxed *Casp8* allele in *Casp8*^{C362S/WT} and *Casp8*^{C362S/WT}*Mlkl*^{-/-} macrophages (*Casp8*^{C362S/-} and *Casp8*^{C362S/-}*Mlkl*^{-/-}) resulted in the formation of ASC specks and the increased secretion of IL-1 β (Fig. 3e). By contrast, ablation of caspase-8 in BMDMs derived from *Casp8*^{WT/WT} mice did not result in the release of IL-1 β or the formation of ASC speck (Extended Data Fig. 6c). Together, these data suggest that the expression of catalytically inactive caspase-8 is required and sufficient to induce inflammasome formation when the active wild-type *Casp8* gene is ablated. Notably, inhibition of caspase activity in BMDMs using IDN-6556 did not recapitulate the data obtained by the expression of CASP8(C362S). Consistent with previous findings¹¹, IDN-6556 was only able to induce the release of IL-1 β and formation of ASC specks when BMDMs were co-treated with lipopolysaccharide (LPS) (Extended Data Fig. 6d). In contrast to CASP8(C362S) expression, the release of IL-1 β induced by IDN-6556 and LPS treatment was completely abolished in cells that lacked MLKL.

To assess whether caspase-1 or ASC contributes to the pathology that caused the lethality found in *Casp8*^{C362S/C362S}*Mlkl*^{-/-} mice, we bred these mice into caspase-1 knockout²¹ or ASC knockout²² genetic backgrounds (Extended Data Fig. 7a). Indeed, *Casp8*^{C362S/C362S}*Mlkl*^{-/-}*Casp1*^{-/-} and *Casp8*^{C362S/C362S}*Mlkl*^{-/-}*Asc*^{-/-} mice survived beyond parturition

(beyond 20 weeks of age) and developed normally without any major macroscopic alterations (Fig. 4a). Both genotypes led to abnormal haematopoiesis, which was characterized by a strong increase in spleen size and weight (Fig. 4a and Extended Data Fig. 7b), as observed in *Casp8^{-/-} Mkl1^{-/-}* mice⁶. Histological analyses did not display overwhelming tissue damage or loss of Paneth cells in the intestines of *Casp8^{C362S/C362S} Mkl1^{-/-} Asc^{-/-}* or *Casp8^{C362S/C362S} Mkl1^{-/-} Casp1^{-/-}* mice (Fig. 4b, c and Extended Data Fig. 5a), suggesting that the lethality of mice that express catalytically inactive caspase-8 is mainly caused by caspase-1 and ASC. Notably, we still detected the formation of ASC specks and aggregation of caspase-8 in intestinal tissues of *Casp8^{C362S/C362S} Mkl1^{-/-} Casp1^{-/-}* mice (Fig. 4b, d and Extended Data Fig. 7d).

Our data collectively demonstrate that catalytically inactive caspase-8 serves as a nucleation signal for the formation of ASC specks and activation of caspase-1, which ultimately leads to the premature death of mice when necroptosis is blocked. These results reveal a previously unknown and unexpected role for the enzymatic activity and scaffold function of caspase-8, which involves the activation of the inflammasome and induction of pyroptosis under circumstances in which apoptosis and necroptosis are compromised. Notably, we found that the inhibition of necroptosis alone was sufficient to prevent embryonic lethality when CASP8(C362S) was specifically expressed in the endothelial or skin compartment (Fig. 2). In contrast to IECs or macrophages, endothelial cells and skin epithelium that expressed CASP8(C362S) did not undergo pyroptosis (Fig. 2 and Extended Data Fig. 7c), indicating that the capability of the caspase-8 scaffold to induce pyroptosis is restricted to specific cell types—such as myeloid cells and IECs—that need to respond regularly to invading microbial pathogens. Caspase-8 has frequently been reported to interact with the caspase-1–ASC adaptor complex and to promote ASC self-assembly, particularly during bacterial infection^{23–27}. Viruses are heavily reliant on the fate of infected cells and have evolved to encode suppressors of apoptosis that inhibit caspase-8 and necroptosis suppressors that inhibit RHIM-containing proteins, such as RIPK1 and RIPK3²⁸. We therefore hypothesize that the abundance of such viral inhibitors may have driven the counteradaptation of pyroptosis as a host defence. Thus, the caspase-8-mediated switch between different modes of cell death adds a critical layer to the plasticity of specific pathogen-tailored immune responses. Questions that remain to be addressed include how the different modes of inflammatory and/or lytic cell death after inhibition of the enzymatic activity of caspase-8 influence anti-microbial immunity and coordinate adaptive immune responses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1770-6>.

1. Muzio, M. et al. FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death-inducing signaling complex. *Cell* **85**, 817–827 (1996).

2. Boldin, M. P., Goncharov, T. M., Goltsev, Y. V. & Wallach, D. Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1- and TNF receptor-induced cell death. *Cell* **85**, 803–815 (1996).
3. Varfolomeev, E. E. et al. Targeted disruption of the mouse Caspase 8 gene ablates cell death induction by the TNF receptors, Fas/Apo1, and DR3 and is lethal prenatally. *Immunity* **9**, 267–276 (1998).
4. Kaiser, W. J. et al. RIP3 mediates the embryonic lethality of caspase-8-deficient mice. *Nature* **471**, 368–372 (2011).
5. Oberst, A. et al. Catalytic activity of the caspase-8-FLIP_L complex inhibits RIPK3-dependent necrosis. *Nature* **471**, 363–367 (2011).
6. Alvarez-Diaz, S. et al. The pseudokinase MLKL and the kinase RIPK3 have distinct roles in autoimmune disease caused by loss of death-receptor-induced apoptosis. *Immunity* **45**, 513–526 (2016).
7. Kang, T. B. et al. Caspase-8 serves both apoptotic and nonapoptotic roles. *J. Immunol.* **173**, 2976–2984 (2004).
8. Günther, C. et al. Caspase-8 regulates TNF-α-induced epithelial necroptosis and terminal ileitis. *Nature* **477**, 335–339 (2011).
9. Hartwig, T. et al. The TRAIL-induced cancer secretome promotes a tumor-supportive immune microenvironment via CCR2. *Mol. Cell* **65**, 730–742 (2017).
10. Henry, C. M. & Martin, S. J. Caspase-8 acts in a non-enzymatic role as a scaffold for assembly of a pro-inflammatory “FADDosome” complex upon TRAIL stimulation. *Mol. Cell* **65**, 715–729 (2017).
11. Kang, S. et al. Caspase-8 scaffolding function and MLKL regulate NLRP3 inflammasome activation downstream of TLR3. *Nat. Commun.* **6**, 7515 (2015).
12. Philip, N. H. et al. Activity of uncleaved Caspase-8 controls anti-bacterial immune defense and TLR-induced cytokine production independent of cell death. *PLoS Pathog.* **12**, e1005910 (2016).
13. Su, H. et al. Requirement for caspase-8 in NF-κB activation by antigen receptor. *Science* **307**, 1465–1468 (2005).
14. Constien, R. et al. Characterization of a novel EGFP reporter mouse to monitor Cre recombination as demonstrated by a Tie2 Cre mouse line. *Genesis* **30**, 36–44 (2001).
15. Hafner, M. et al. Keratin 14 Cre transgenic mice authenticate keratin 14 as an oocyte-expressed protein. *Genesis* **38**, 176–181 (2004).
16. Madison, B. B. et al. *cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277**, 33275–33283 (2002).
17. Kovalenko, A. et al. Caspase-8 deficiency in epidermal keratinocytes triggers an inflammatory skin disease. *J. Exp. Med.* **206**, 2161–2177 (2009).
18. Peitz, M., Pfannkuche, K., Rajewsky, K. & Edenhofer, F. Ability of the hydrophobic FGF and basic TAT peptides to promote cellular uptake of recombinant Cre recombinase: a tool for efficient genetic engineering of mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 4489–4494 (2002).
19. Dannappel, M. et al. RIPK1 maintains epithelial homeostasis by inhibiting apoptosis and necroptosis. *Nature* **513**, 90–94 (2014).
20. Broz, P. & Dixit, V. M. Inflammasomes: mechanism of assembly, regulation and signalling. *Nat. Rev. Immunol.* **16**, 407–420 (2016).
21. Van Gorp, H. et al. Familial Mediterranean fever mutations lift the obligatory requirement for microtubules in Pyrin inflammasome activation. *Proc. Natl Acad. Sci. USA* **113**, 14384–14389 (2016).
22. Drexler, S. K. et al. Tissue-specific opposing functions of the inflammasome adaptor ASC in the regulation of epithelial skin carcinogenesis. *Proc. Natl Acad. Sci. USA* **109**, 18384–18389 (2012).
23. Chen, M. et al. Internalized *Cryptococcus neoformans* activates the canonical caspase-1 and the noncanonical caspase-8 inflammasomes. *J. Immunol.* **195**, 4962–4972 (2015).
24. Man, S. M. et al. *Salmonella* infection induces recruitment of caspase-8 to the inflammasome to modulate IL-1β production. *J. Immunol.* **191**, 5239–5246 (2013).
25. Pierini, R. et al. AIM2/ASC triggers caspase-8-dependent apoptosis in *Francisella*-infected caspase-1-deficient macrophages. *Cell Death Differ.* **19**, 1709–1721 (2012).
26. Van Oudenbosch, N. et al. Caspase-1 engagement and TLR-induced c-FLIP expression suppress ASC/Caspase-8-dependent apoptosis by inflammasome sensors NLRP1b and NLRC4. *Cell Rep.* **21**, 3427–3444 (2017).
27. Gurung, P. et al. FADD and caspase-8 mediate priming and activation of the canonical and noncanonical Nlrp3 inflammasomes. *J. Immunol.* **192**, 1835–1846 (2014).
28. Mocarski, E. S., Upton, J. W. & Kaiser, W. J. Viral infection and the evolution of caspase 8-regulated apoptotic and necrotic death pathways. *Nat. Rev. Immunol.* **12**, 79–88 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

Mice

Casp8^{C362S} mice were generated by pronuclear injection of C57BL/6N zygotes with 20 ng μl^{-1} *Cas9* mRNA (TriLink Biotechnologies), 10 ng μl^{-1} sgRNA and 20 ng μl^{-1} template DNA (Eurofins). Sequence of single-guide RNA (5'-CACCGTTTCATTCAGGCTTGCCA-3') and template DNA (5'-CACTGGTTCAAAGTGCCTTCCCTGTCTGGAAACCCAAGATCTTTTCATTCAGGCTAGCCAAGGAAGTAAGTCCAGAAAGGAGTGCCTGATGAGGCAGGCTTCGAGCAACAGAAC-3'). Ear cuts were genotyped by Sanger sequencing (Microsynth SEQLAB) with PCR Primer (5'-TGCAAATGAAATCCACGAGA-3' and 5'-CCAGGTTCCATTCACAGGAT-3'). Founders carrying the intended C362S mutation were backcrossed to a C57BL/6N genetic background for five generations. All mouse studies were performed after approval by local government authorities (LANUV, NRW, Germany) in accordance with the German animal protection law. Animals were housed in the animal care facility of the University of Cologne under standard pathogen-free conditions with a 12-h light/dark schedule and provided with food and water ad libitum. Calculations to determine sample size, randomization and blinding were not performed. Mice were grouped according to their phenotype in mixed sexes.

Embryological studies

For timed mating experiments a male mouse was paired with a single female mouse. Embryonic day count started at E0.5 with the day at which a positive plug was found. For staining of vascularization of the yolk sac, the yolk sac was fixed in 4% PFA in PBS for 1 h at 4 °C, washed in PBS with 0.5% Tween-20 followed by blocking with blocking buffer (PBS with 0.5% Tween-20, 0.2% BSA and 2% normal goat serum) for 2 h. Yolk sacs were then incubated with Alexa Fluor 647 anti-mouse CD31 antibody (BioLegend) overnight at 4 °C, washed with PBS and mounted with Mowiol. Imaging was conducted on a motorized inverted Olympus IX81 microscope (Cell Imaging Software)²⁹.

Immunohistochemistry

Intestinal tissues and skin were fixed in Roti-HistoFix 4% (Carl Roth), embedded in paraffin and cut in 5- μm sections³⁰. After rehydration and heat-induced antigen retrieval in 10 mM citrate buffer or by proteinase K treatment, sections were stained with antibodies against ASC (Santa Cruz), caspase-8 (Enzo) and lysozyme (DAKO). As secondary antibodies, biotinylated goat anti-rabbit IgG (Perkin Elmer) and biotin-SP-AffiniPure goat anti-rat IgG (Jackson ImmunoResearch) were used. Staining was visualized using the ABC Kit Vectastain Elite (Vector Laboratories) and DAB staining kit (DAKO) and counterstained with haematoxylin (Carl Roth). Incubation time of the substrate for DAB staining was equal for all tissue sections. Tissue sections were stained with H&E. Goblet cells were stained using PAS staining (Sigma Aldrich) according to the manufacturer's instructions. Stained sections were scanned with an SCN4000 Slide Scanner (Leica) and analysed with the imaging software Aperio ImageScope v.12.2.2.5015 (Leica). For Paneth cell counts, we analysed 30 crypts per mouse and for dead-cell counts we analysed 15 villi per mouse.

Immunofluorescence microscopy

Immunofluorescence microscopy analysis was carried out as described previously³⁰. In brief, cells were seeded on glass coverslips, treated as indicated and fixed with 3% PFA in PBS for 20 min at room temperature. Subsequently, cells were washed twice with PBS and incubated with blocking buffer (0.1% saponin (Carl Roth), 3% BSA (Carl Roth) in PBS) for 30 min at room temperature. Coverslips were incubated with diluted primary anti-ASC antibody (Santa Cruz), or human-specific anti-caspase-8 antibody (Cell Signaling) in blocking buffer in a humid chamber overnight at 4 °C. After incubation, coverslips were washed with washing buffer (0.1% saponin in PBS) three times and incubated

with secondary antibody goat anti-rabbit Alexa Fluor 568 (Thermo Fisher Scientific) or goat anti-mouse Alexa Fluor 488 (Thermo Fisher Scientific) for 1 h at room temperature. Subsequently, cells were stained with 300 nM DAPI (Molecular Probes) for 10 min and washed three times, before being embedded with Mowiol overnight. Imaging was performed on an UltraView Vox Spinning Disk confocal microscope (Perkin Elmer and Nikon) and analysed using Volocity v.5.4.2 (PerkinElmer).

Endothelial cell culture

Mouse endothelial cells were isolated from the lungs of *Casp8*^{WT/WT} and *Casp8*^{C362S/WT} mice. Organs were resected and briefly washed in PBS, before being minced and enzymatically (0.5% collagenase) digested. The cell solution was then squeezed through a cell strainer (70 μm) and processed for magnetic bead separation (mouse CD31, Miltenyi Biotec) according to the manufacturer's protocol. CD31⁺ endothelial cells were seeded on gelatine-coated wells and cultured in a 1:1 mixture of EGM2 (PromoCell) and full supplemented Dulbecco's modified Eagle medium (DMEM) (Merck) (containing 20% FCS, 4 g l⁻¹ glucose, 2 mM glutamine, 1% penicillin-streptomycin (100 U ml⁻¹ penicillin, 100 μg ml⁻¹ streptomycin), sodium pyruvate 1% (1 mM), HEPES (20 mM) and 1% non-essential amino acids). After the first passage, cells were resorted using the same magnetic beads. Isolated primary endothelial cells were analysed using the anti-CD31 antibody to distinguish endothelial cells from other cell types and routinely tested negative for mycoplasma contamination by PCR. DNA fragments of *Casp8*^{WT/WT} and *Casp8*^{C362S/WT} alleles were excised by treatment with recombinant HTNCre (5 μM) purified from *Escherichia coli* for 24 h in a mixture of DMEM:PBS 1:1 twice. Complete knockout was confirmed by PCR and western blot analysis. *Casp8*^{-/-} cells were used as controls. Endothelial cells were treated with mouse TNF (R&D), cycloheximide (Sigma) and Nec-1 (Enzo) as indicated.

Purification of HTNCre

For site-specific recombination of floxed *Casp8* alleles, HTNCre from transformed *E. coli* was purified. In brief, bacteria were grown from a diluted overnight culture until an optical density at 600 nm of 0.6–1.1 was reached. Transcription of the HTNCre construct was initiated by the addition of IPTG (1 mM). The culture was incubated for additional 4 h at 37 °C and then centrifuged for 25 min at 8,500 rpm at 4 °C. The pellet (1 g) was resuspended in 10 ml PBS, 1 mg lysozyme per ml suspension, 1:1,000 benzonase (Novagen) and protease inhibitor (Roche). Lysate was homogenated through a high-pressure homogenizer and HTNCre was purified by HisTrap FF crude column (GE Healthcare).

Macrophage differentiation

Mouse BMDMs were differentiated from the bone marrow of mice with the indicated genotypes. BMDMs were generated by culturing mouse bone marrow in RPMI supplemented with 15% L929-conditioned medium, 10% FCS (Sigma), 10 mM HEPES (Biochrom), 1 mM sodium pyruvate (Biochrom), 2 mM L-glutamine (Biochrom), 100 U ml⁻¹ penicillin, 100 μg ml⁻¹ streptomycin for 7 days. DNA fragments of *Casp8*^{WT/WT}, *Casp8*^{C362S/WT} and *Casp8*^{fl/fl} alleles were excised by treatment with recombinant HTNCre (2.5 μM) (Excellgene, purity of >98%, endotoxin levels of <0.1 endotoxin units μg^{-1}) for 24 h in a mixture of RPMI:PBS 1:1 (v/v) or LPS (200 ng ml⁻¹) (Invivogen) and IDN-6556 (20 μM).

Viability assay

Viability was detected using the neutral red assay. In brief, 20,000 endothelial cells per well were seeded on gelatine-coated 96-well plates and cultured in the appropriate medium overnight. Cells were then exposed to the specific conditions and neutral red assay was performed 6 h after treatment.

Cell death assay

LDH release was measured to analyse cell death using the cytotoxicity detection kit (Roche) according to the manufacturer's instructions.

Tissue homogenates

Tissue sections of the small intestine from mice with the indicated genotypes were homogenized in RIPA buffer containing protease and phosphatase inhibitors (Roche) (20 ml buffer per 1 g tissue) using gentleMACS C tubes (Miltenyi Biotec). Supernatants of tissue homogenates were used for the determination of cytokine levels.

Blood parameter analysis

After cervical dislocation of mice, blood was collected from the heart with an EDTA-coated syringe and immediately diluted with Cellpack (Sysmex) in a ratio of 1:5. Blood was analysed at the University Hospital Cologne, Institute for Clinical Chemistry.

Cytokine ELISA

Cytokine levels were determined using IL-1 β ELISA (R&D Systems) according to the manufacturer's instructions.

Cytokine array

The LUNARIS Mouse Cytokine Kit (AYOXXA Biosystems) was used to determine cytokine levels in tissue homogenates according to the manufacturer's instructions.

Cultivation and transfection of human cells

HCT-116 and HEK293T cells were purchased from ATCC. All cell lines routinely tested negative for Mycoplasma contamination by PCR. HCT-116 cells were cultured in McCoy's 5A modified medium (Merck) supplemented with 10% heat-inactivated FCS (Biowest) and transfected with Lipofectamine 2000 (Invitrogen). HEK293T cells were cultured in DMEM (Merck) supplemented with 10% heat-inactivated FCS (Biowest) and were transfected with polyethylenimine (Polysciences Europe GmbH). HEK293T and HCT-116 cells were transfected for 14 h with respective constructs and subsequently treated with IDN-6556 (MedChemExpress) and Nec-1 (Enzo) as indicated. The coding sequence of *CASP8* and site-directed *CASP8*^{C360S} mutation (forward, 5'-ATTCAGGCTTCTCAGGGGAT-3'; reverse, 5'-ATCCCCCTGAGAAGCCTGAAT-3') (Eurofins) was cloned into pcDNA3.1+. The coding sequence of human *ASC* was cloned into pDsRed2.

Generation of CRISPR-Cas9 *Casp8* knockout cells

Oligonucleotide sgRNAs (1, GCTCTTCCGAATTAATAGAC; 2, CTACCTAAACACTAGAAAGG) (Eurofins) targeting the *Casp8* locus were cloned into the pSpCas9(BB)-2A-GFP (PX458) vector, which was a gift from F. Zhang (Addgene plasmid 48138), and transfected into HCT-116 and HEK293T cells. After transfection for 24 h, cells were plated onto 96-well plates (1 cell per well) and caspase-8 deficiency was checked after 3–4 weeks in single-cell clones by western blot analysis.

Immunoprecipitation

HEK293T cells were transfected for 14 h with respective constructs and subsequently treated with IDN-6556 as indicated. Cells were trypsinized, washed twice with chilled PBS and centrifuged at 700g for 3 min at 4 °C. The cell pellet was resuspended in RIPA buffer (1% Triton X-100, 150 mM NaCl, 50 mM Tris, 0.1% SDS, 0.5% deoxycholate, 10% glycerol and 1 mM EGTA) and incubated for 30 min on ice and centrifuged for 20 min at 20,800g at 4 °C. Before antibody incubation, immunoprecipitation incubation buffer (20 mM Tris, pH 8.0, 137 mM NaCl, 1% NP-40 and 2 mM EDTA) was added in a ratio of 1 \times RIPA lysate:2 \times incubation buffer. Lysates were incubated with an anti-caspase-8 human-specific antibody (Cell Signaling) for 1 h at 4 °C before addition of μ MACS Protein G MicroBeads (Miltenyi Biotec). Normal mouse IgG (Santa Cruz) served as a control. Immunoprecipitation was performed according to the manufacturer's instructions.

Western blot analysis

Cells and tissue were lysed in 20 mM Tris-HCl pH 7.5, 135 mM NaCl, 1.5 mM MgCl₂, 1 mM EGTA, 1% Triton X-100, 10% glycerol, protease

inhibitor (Roche) and phosphatase inhibitor (Roche). After 20 min on ice, cells were centrifuged at 20,000g for 20 min at 4 °C, the soluble fraction was collected and the insoluble fraction was mechanically disrupted in 6 M urea, 3% SDS, 10% glycerine and 50 mM Tris pH 6.8. Cell lysates of endothelial, HCT-166 and HEK293T cells were prepared in CHAPS lysis buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 1% CHAPS, protease inhibitor (Roche) and phosphatase inhibitor (Roche))³¹. Protein concentrations of cell lysates and tissue lysates were determined using a Pierce BCA Protein Assay Kit (ThermoFisher Scientific) or DC protein assay (Bio Rad) according to the manufacturer's instructions. Proteins were separated by SDS-PAGE and transferred to a nitrocellulose or PVDF membrane. Proteins were stained with antibodies against ASC (Santa Cruz), caspase-1 p10 (Santa Cruz), caspase-1 p20 (Adipogen), caspase-1 mouse-specific (Biolegend), caspase-8 (Enzo), caspase-8 mouse-specific (Cell Signaling), caspase-8 human-specific (Cell Signaling), cleaved caspase-8 Asp387 (Cell Signaling), caspase-3 (Cell Signaling), cleaved caspase-3 (Cell Signaling), phosphorylated MLKL (S345) (Abcam), DsRed (BD Biosciences), human-specific caspase-7 (Cell Signaling), human-specific caspase-9 (Cell Signaling), FADD (BD Biosciences), RIPK1 (BD Biosciences), RIPK3 (Enzo), cFLIP (Sigma) and β -actin HRP-conjugated (Santa Cruz), secondary antibodies included goat anti-rabbit IgG conjugated to horseradish peroxidase (HRP, Cell Signaling), goat anti-mouse IgG HRP (Sigma), goat anti-mouse IgG light chain HRP (Jackson Immuno Research) and goat anti-rat IgG (H+L) HRP (ThermoFisher Scientific) and then developed using a ChemiDoc MP Imaging System (BioRad)³⁰.

Statistics

Data are mean \pm s.e.m. Sample sizes (replicates, animals) are traceable as individual data points in each figure. In vitro experiments were repeated at least twice. Data involving animals depict pooled data of at least two independent experiments. All statistical tests used to examine statistical significance were two-sided. Exact *P* values and the respective tests or analyses are listed in the figure legends; **P* < 0.05, ***P* < 0.01, ****P* < 0.001; NS, not significant. LUNARIS Analysis Suite 1.3, GraphPad Prism 7.0 and Excel were used to analyse data in this study.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data supporting the findings of this study are available within the paper and its Supplementary Information. Source Data for Figs. 1–4 and Extended Data Figs. 3, 5–7 are provided with the paper.

29. Witt, A. et al. IAP antagonization promotes inflammatory destruction of vascular endothelium. *EMBO Rep.* **16**, 719–727 (2015).
30. Andree, M. et al. BID-dependent release of mitochondrial SMAC dampens XIAP-mediated immunity against *Shigella*. *EMBO J.* **33**, 2171–2187 (2014).
31. Kashkar, H. et al. XIAP-mediated caspase inhibition in Hodgkin's lymphoma-derived B cells. *J. Exp. Med.* **198**, 341–347 (2003).

Acknowledgements We thank M. Menning, A. Manav, T. Roth and R. Hoppe for technical assistance; the CECAD in vivo Research Facility and their Transgenic Core Unit for mouse care and the generation of transgenic mice (B. Zevnik); Imaging Facilities of the CECAD and the Collaborative Research Center 670 (SFB670, Z2); T. Wunderlich for the HTNCre-expressing bacterial strain; G. Malchau for supporting blood analyses; S. Hedrick for the *Casp8*^{fl/fl} mouse and J. Tschopp for the *Asc*^{-/-} mouse. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) CRC670, CRC1218, CRU286 and The German Cancer Aid to H.K. and by the ERC (grant agreements 323040 and 787826) to M.P.

Author contributions M.F. generated the *Casp8*^{C362S} mice, performed genetic crosses and carried out most of the experimental work. S.D.G. performed microscopy analysis, BMDM experiments, mouse preparations and designed the figures. R.S. was involved in designing the

Article

knock-in strategy for Casp8^{C362S} mice and carried out the histopathology analysis of intestines. M.-C.A. performed immunoprecipitation experiments, endothelial cell experiments and mouse preparations. F.S. carried out overexpression studies, generated knockout cell lines and performed BMDM experiments. J.P.W. performed BMDM experiments and statistical analysis of intestinal inflammation and pyroptosis. L.M.S. and N.S. isolated and carried out endothelial cell work. H.S. evaluated knockout cell lines. J.M.S. supported mouse work. M.K. provided essential reagents. M.L. provided essential mouse lines. M.P. provided essential mouse lines and was involved in the design of the study. H.K. designed and supervised the study. All authors analysed the data, discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

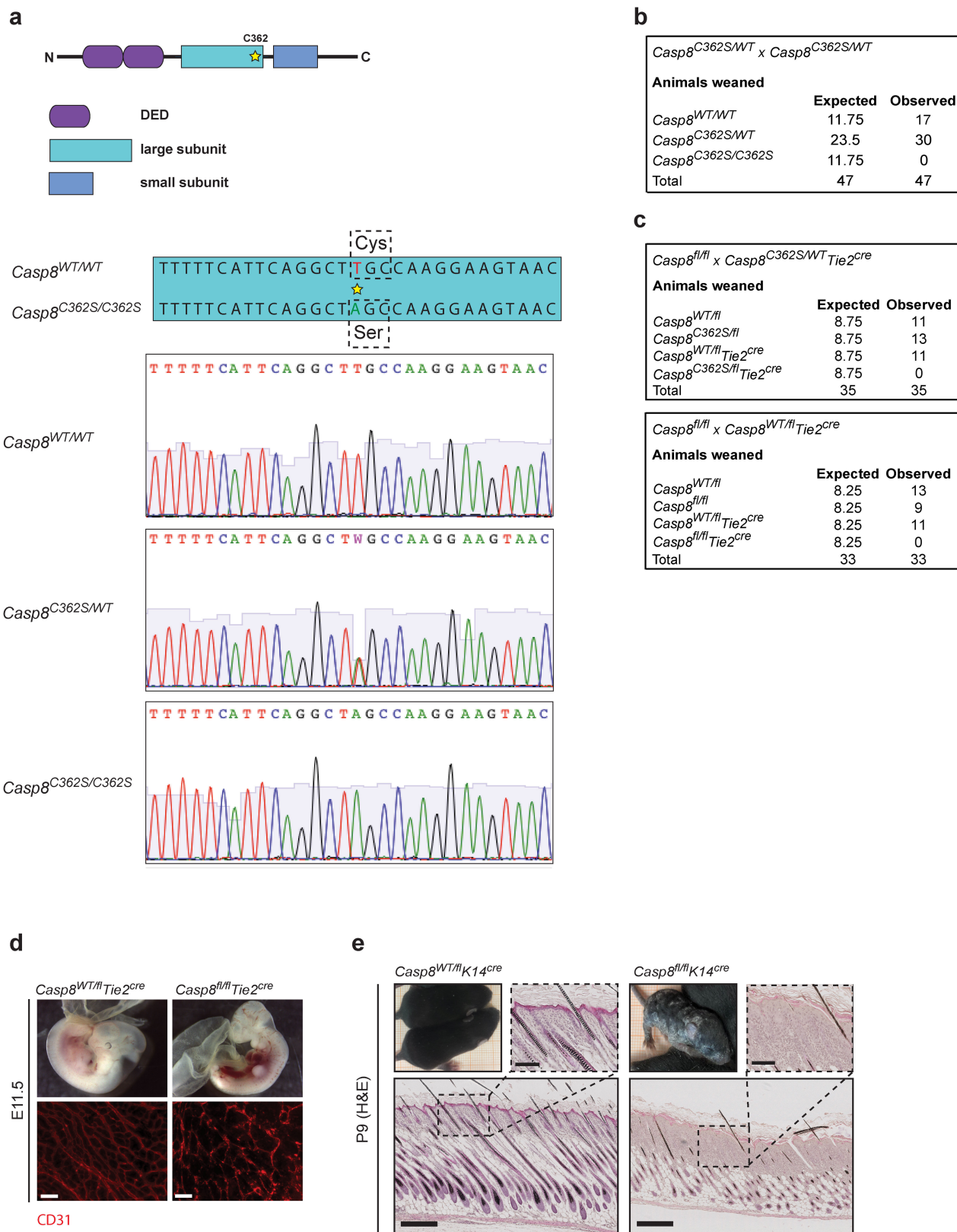
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1770-6>.

Correspondence and requests for materials should be addressed to H.K.

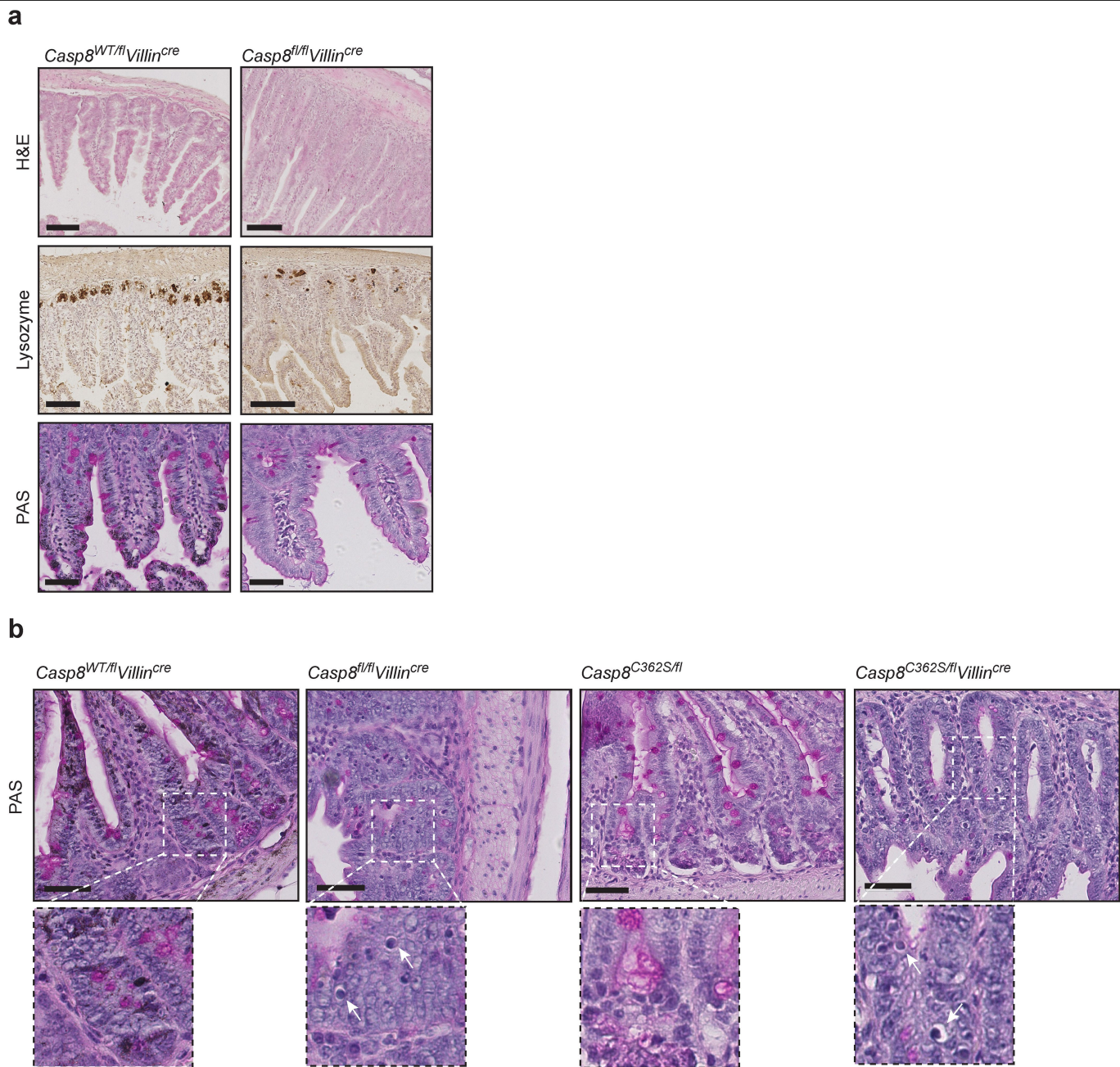
Peer review information *Nature* thanks Igor E. Brodsky, William Kaiser and Seamus Martin for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



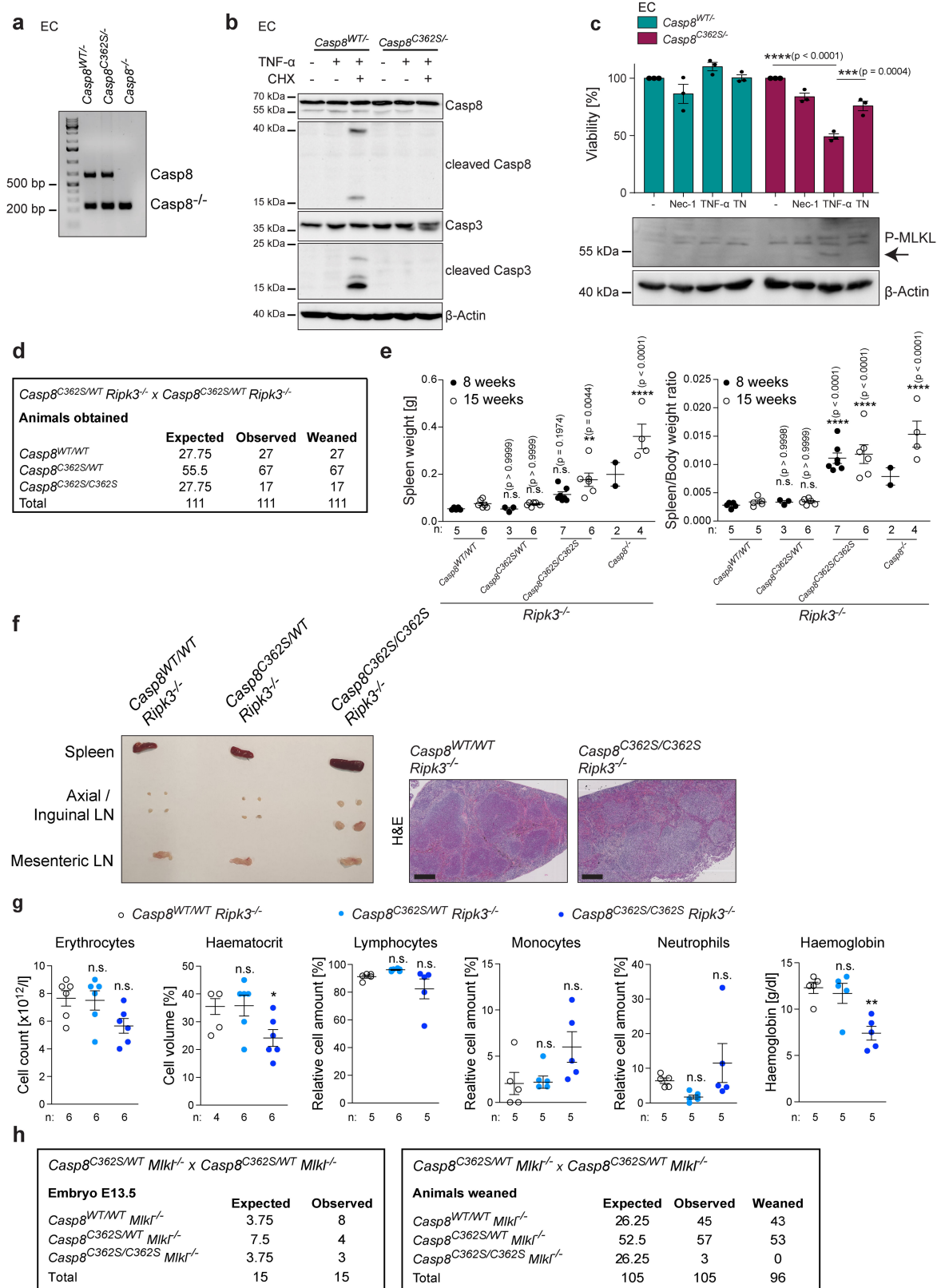
Extended Data Fig. 1 | The enzymatic activity of caspase-8 is required to inhibit necroptosis. **a**, Top, schematic illustration of the *Casp8* gene with the domain structure and the position of catalytic cysteine (star). Bottom, targeted genomic sequence of *Casp8* and representative sequence analysis of embryos at E11.5 with respective genotypes. **b**, **c**, Expected and observed numbers of mice per genotype obtained from the indicated crossings. **d**, Representative

images of *Casp8*^{WT/fl} *Tie2*^{cre} ($n = 2$) and *Casp8*^{fl/fl} *Tie2*^{cre} ($n = 5$) mouse embryos at E11.5 (top). Whole-mount yolk sacs stained with anti-CD31 antibody as endothelial marker (bottom). Scale bars, 100 μ m. **e**, Representative images of 9-day-old *Casp8*^{WT/fl} *K14*^{cre} ($n = 5$) and *Casp8*^{fl/fl} *K14*^{cre} ($n = 4$) mice (top left) and skin sections stained with H&E (top right and bottom). Scale bars, 100 μ m (magnification, top) and 300 μ m (bottom).



Extended Data Fig. 2 | The enzymatic activity of caspase-8 is required to inhibit necroptosis. a, Representative images of ileal sections from 10-week-old *Casp8^{WT/fi}Villin^{cre}* ($n=4$) and *Casp8^{fl/fi}Villin^{cre}* ($n=3$) mice stained with H&E (top), immunostained for lysozyme (Paneth cells, middle) and PAS (bottom).

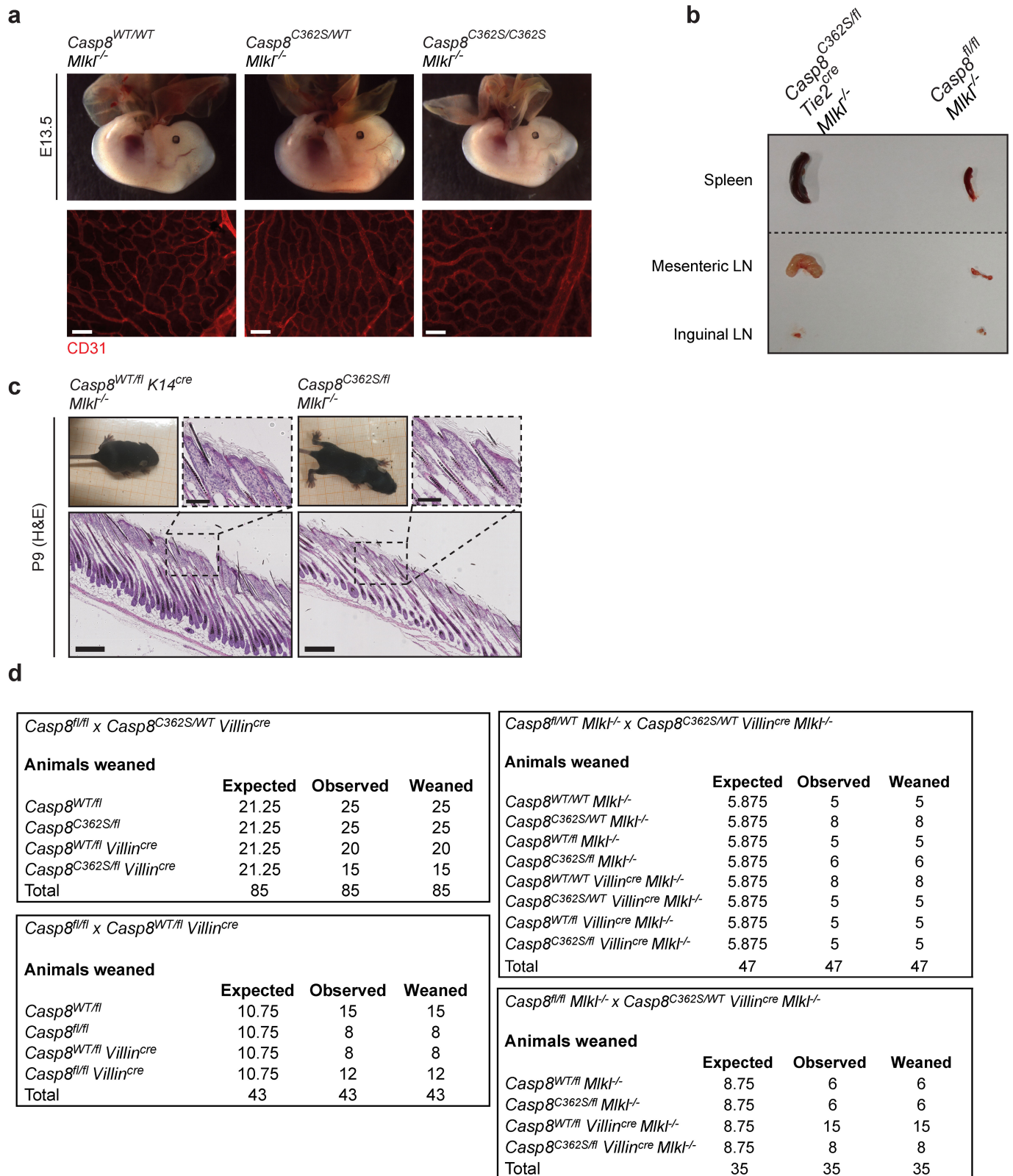
Scale bars, 100 μ m. **b**, Representative images of ileal sections from 10-week-old *Casp8^{WT/fi}Villin^{cre}* ($n=4$), *Casp8^{fl/fi}Villin^{cre}* ($n=3$), *Casp8^{C362S/fi}* ($n=3$) and *Casp8^{C362S/fi}Villin^{cre}* ($n=4$) mice stained with PAS. Arrows, dead cells. Scale bars, 50 μ m.



Extended Data Fig. 3 | See next page for caption.

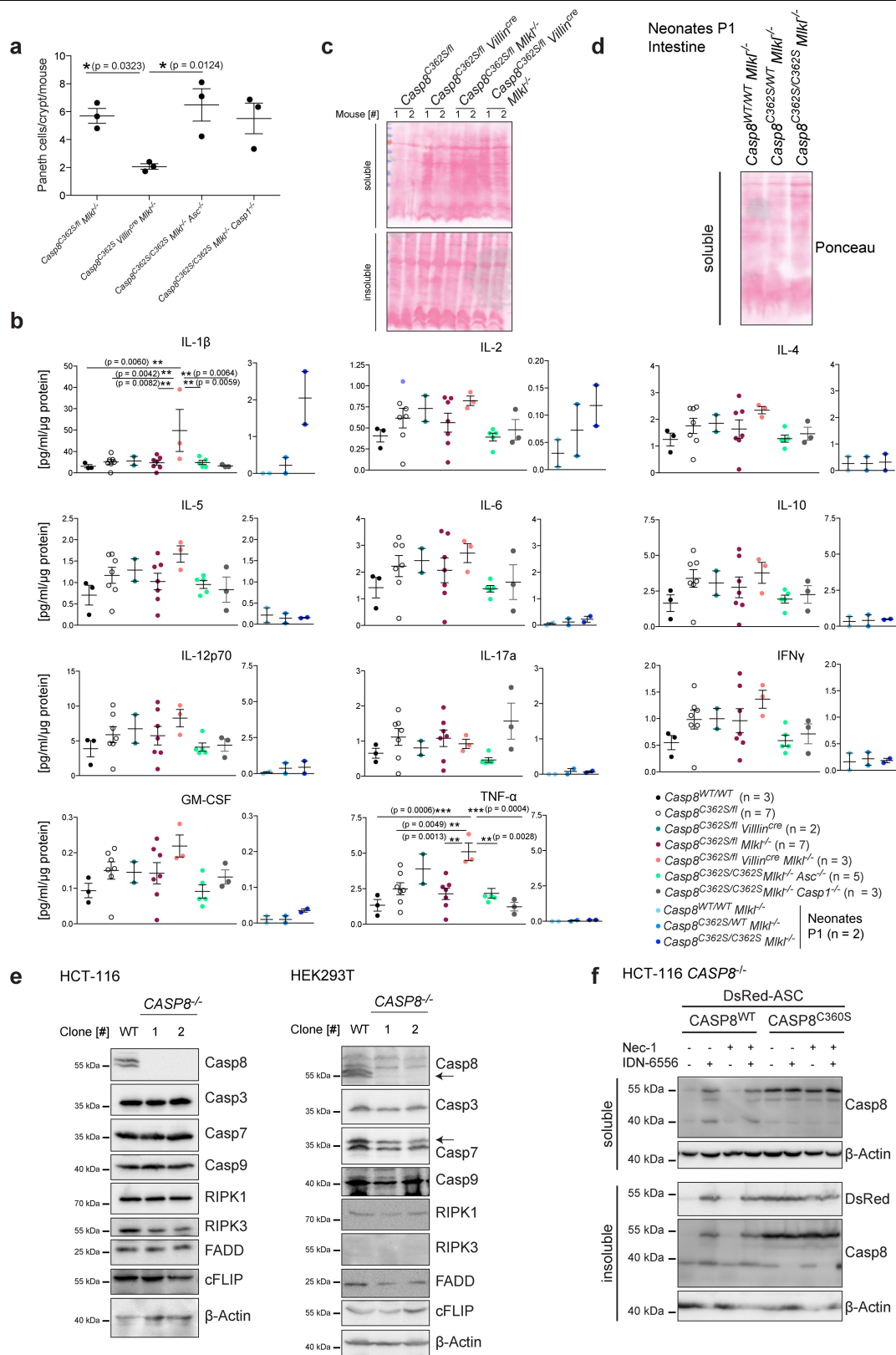
Extended Data Fig. 3 | CASP8(C362S) induces necroptosis-independent tissue destruction. **a**, Genotyping PCR of respective endothelial cells (ECs) after treatment with cell-permeable recombinant HTNCre protein. Results are representative of two individual experiments. **b**, Analysis of caspase-8 and caspase-3 processing by western blot after treatment with TNF (10 ng ml⁻¹), CHX (2.5 µg ml⁻¹) or both (TNF and CHX). Results are representative of two individual experiments. **c**, Top, viability of endothelial cells after treatment with TNF (10 ng ml⁻¹), Nec-1 (30 µM) or both (TNF and Nec-1) for 6 h (top) in biologically independent replicates. Dots represent individual biological replicates ($n = 3$). Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis. $n = 3$, representative of two individual experiments. Bottom, western blot analysis of cell lysates examining phosphorylated MLKL (P-MLKL) and β -actin. Results are representative of two individual experiments. **d**, Expected, observed and weaned numbers of mice per genotype obtained from the indicated crossings. **e**, Spleen weight and spleen:body weight ratios of 8- and 15-week-old mice of the indicated genotypes. Dots and circles, individual

mice. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis compared to the corresponding *Casp8*^{WT/WT} values. **f**, Representative images of spleen, axil and inguinal lymph nodes (LN) and mesenteric lymph nodes (left) as well as splenic sections from *Casp8*^{WT/WT}*Ripk3*^{-/-} ($n = 3$), *Casp8*^{C362S/WT}*Ripk3*^{-/-} ($n = 3$) and *Casp8*^{C362S/C362S}*Ripk3*^{-/-} mice ($n = 3$) stained with H&E (right) from 15-week-old mice. Scale bars, 300 µm. **g**, Cardiac blood was analysed for cell numbers, and haematocrit and haemoglobin concentrations from 8–15-week-old mice. Dots and circles, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Dunnett's post-analysis compared to the corresponding *Casp8*^{WT/WT}*Ripk3*^{-/-} values. Exact P values (from left to right): erythrocytes, $P = 0.9774$, $P = 0.0607$; haematocrit: $P = 0.9961$, $P = 0.0462$; lymphocytes, $P = 0.5812$, $P = 0.2426$; monocytes, $P = 0.9944$, $P = 0.0693$; neutrophils, $P = 0.4614$, $P = 0.4349$; haemoglobin, $P = 0.8326$, $P = 0.0025$. **h**, Expected, observed and weaned numbers of embryos (left) and mice (right) per genotype obtained from the indicated crossings.



Extended Data Fig. 4 | CASP8(C362S) induces necroptosis-independent tissue destruction. **a**, Representative images of *Casp8*^{WT/WT} *Mik1*^{-/-} (*n* = 8), *Casp8*^{C362S/WT} *Mik1*^{-/-} (*n* = 7) and *Casp8*^{C362S/C362S} *Mik1*^{-/-} (*n* = 3) mouse embryos at E13.5 (top). Whole-mount yolk sacs stained with anti-CD31 antibody as endothelial marker (bottom). Scale bars, 100 μ m. **b**, Representative images of spleen, inguinal and mesenteric lymph nodes from *Casp8*^{C362S/fi} *Tie2*^{cre} *Mik1*^{-/-}

(*n* = 2) and *Casp8*^{fi/fi} *Mik1*^{-/-} (*n* = 2) mice. **c**, Representative images of 9-day-old mice (top left) and skin sections stained with H&E (top right, bottom). Scale bars, 100 μ m (magnification, top) and 300 μ m (bottom). **d**, Expected, observed and weaned numbers of mice per genotype obtained from the indicated crossings.



Extended Data Fig. 5 | *CASP8*(C362S) activates the ASC inflammasome in IECs. **a**, Count of Paneth cells (per crypt per mouse, $n = 3$). Dots, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis. **b**, Cytokine array (AYOXXA Lunar) to detect the indicated cytokines in ileal lysates derived from 5-week-old mice (left) or P1 neonates (right). Dots and circles, individual mice. Data are mean \pm s.e.m. One-way ANOVA followed by Turkey's post-analysis. **c**, **d**, Ponceau staining of ileal lysates from 5-week-old mice ($n = 2$) (**c**) and P1 neonates ($n = 1$) (**d**) (Fig. 3b, c). Lanes, individual mice.

e, Western blot analysis of *CASP8*^{WT/WT} and *CASP8*^{-/-} HCT-116 and HEK293T CRISPR-Cas9 cell clones. Results are representative of one individual experiment. **f**, Western blot analysis of the soluble and insoluble fraction of *CASP8*^{-/-} HCT-116 clone 2 with overexpression of either human wild-type caspase-8 or *CASP8*(C360S) after treatment with IDN-6556 (20 μ M), Nec-1 (10 μ M) or both (IDN-6556 and Nec-1) for 14 h. Results are representative of two individual experiments.

Extended Data Fig. 6 | CASP8(C362S) activates the ASC inflammasome in IECs. **a**, Immunofluorescence confocal images of *CASP8*^{-/-} HCT-116 clone 2 overexpressing either human wild-type caspase-8 or CASP8(C360S) together with DsRed-ASC untreated or treated with IDN-6556 (20 μ M) and stained for caspase-8 after 24 h. Scale bar, 20 μ m. Results are representative of two individual experiments. **b**, Immunoprecipitation of *CASP8*^{-/-} HEK293T clone 1 lysates overexpressing either human wild-type caspase-8, CASP8(C360S) or empty vector (-) together with DsRed-ASC untreated or treated with IDN-6556 (20 μ M) as indicated. Results are representative of two individual experiments. **c**, Immunofluorescence confocal images of BMDMs derived from *Casp8*^{fl/fl} or *Casp8*^{C362S/fl} mice treated with HTNCre for 24 h and stained with an anti-ASC antibody (top) or measurement of IL-1 β levels in the supernatant of BMDMs (bottom; $n = 3$ biologically independent replicates). Scale bar, 20 μ m. Dots and

circles represent individual biological replicates. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis compared to the corresponding untreated value. Results are representative of two individual experiments. **d**, ASC-speck-positive BMDMs (top; $n = 100$ of one representative experiment), measurements of IL-1 β levels (middle; $n = 3$ biologically independent replicates) and LDH release (bottom; $n = 3$ biologically independent replicates) in the supernatants of BMDMs after treatment with LPS (200 ng ml⁻¹), IDN-6556 (20 μ M) or both (LPS and IDN-6556) for 24 h. Dots and circles represent individual biological replicates. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis compared to the corresponding untreated value and shown for $P > 0.1$. Results are representative of two individual experiments.

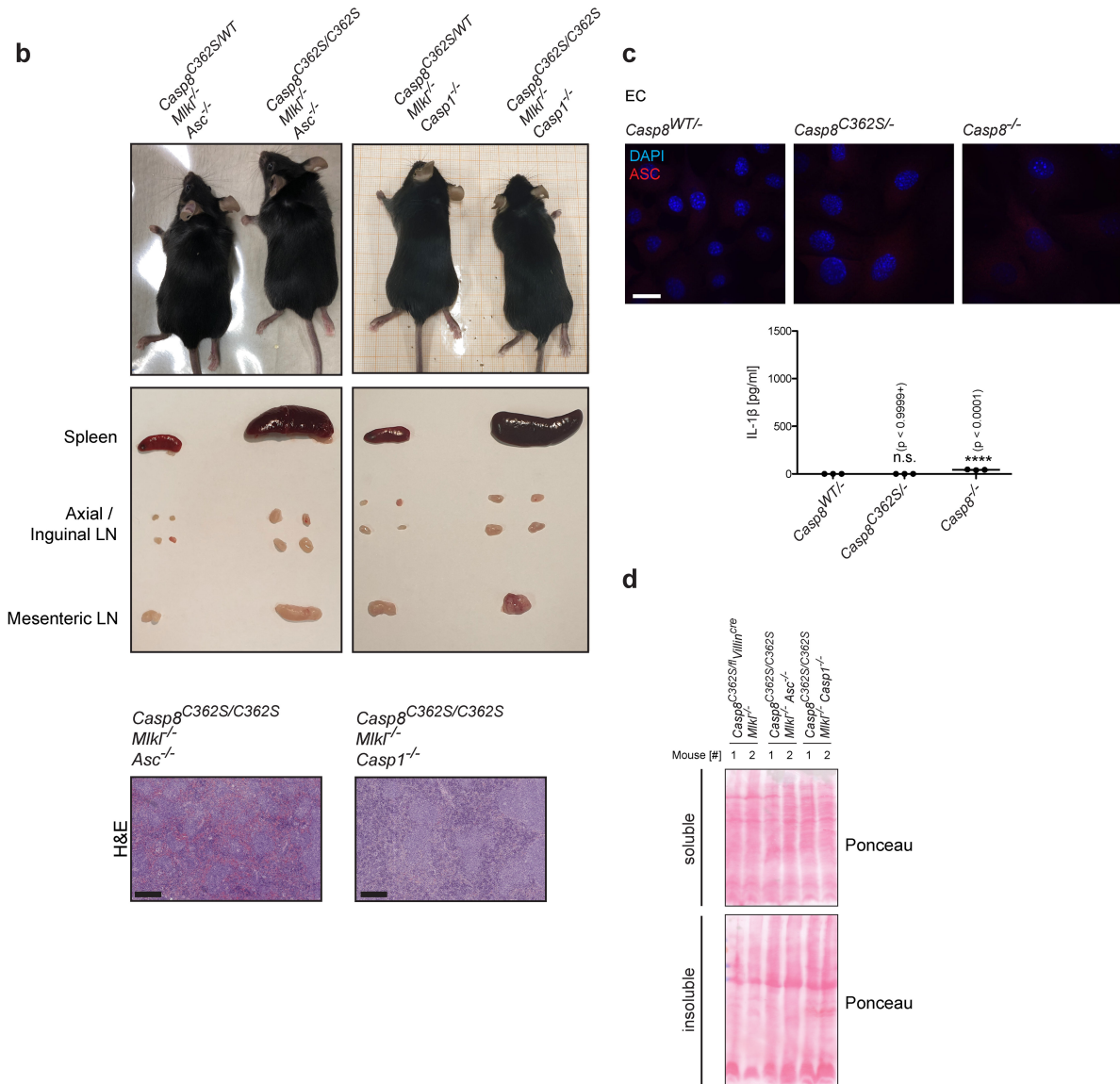
a

<i>Casp8^{C362S/WT} Mik1^{-/-} Asc^{-/-} x Casp8^{C362S/WT} Mik1^{WT/-} Asc^{-/-}</i>			
Animals weaned	Expected	Observed	Weaned
<i>Casp8^{WT/WT} Mik1^{WT/-}</i>	4.5	5	5
<i>Casp8^{C362S/WT} Mik1^{WT/-}</i>	9	14	14
<i>Casp8^{C362S/C362S} Mik1^{WT/-}</i>	4.5	0	0
<i>Casp8^{WT/WT} Mik1^{-/-}</i>	4.5	4	4
<i>Casp8^{C362S/WT} Mik1^{-/-}</i>	9	9	9
<i>Casp8^{C362S/C362S} Mik1^{-/-}</i>	4.5	4	4
Total	36	36	36

<i>Casp8^{C362S/WT} Mik1^{-/-} Asc^{-/-} x Casp8^{C362S/WT} Mik1^{WT/-} Asc^{-/-}</i>			
Animals weaned	Expected	Observed	Weaned
<i>Casp8^{WT/WT}</i>	2.5	2	2
<i>Casp8^{C362S/WT}</i>	5	4	4
<i>Casp8^{C362S/C362S}</i>	2.5	4	4
Total	10	10	10

<i>Casp8^{C362S/WT} Mik1^{-/-} Casp1^{-/-} x Casp8^{C362S/WT} Mik1^{WT/-} Casp1^{-/-}</i>			
Animals weaned	Expected	Observed	Weaned
<i>Casp8^{WT/WT} Mik1^{WT/-}</i>	5.25	6	6
<i>Casp8^{C362S/WT} Mik1^{WT/-}</i>	10.5	12	11
<i>Casp8^{C362S/C362S} Mik1^{WT/-}</i>	5.25	0	0
<i>Casp8^{WT/WT} Mik1^{-/-}</i>	5.25	11	11
<i>Casp8^{C362S/WT} Mik1^{-/-}</i>	10.5	9	9
<i>Casp8^{C362S/C362S} Mik1^{-/-}</i>	5.25	4	3
Total	42	42	40

<i>Casp8^{C362S/WT} Mik1^{-/-} Casp1^{-/-} x Casp8^{C362S/WT} Mik1^{WT/-} Casp1^{-/-}</i>			
Animals weaned	Expected	Observed	Weaned
<i>Casp8^{WT/WT}</i>	3.25	2	2
<i>Casp8^{C362S/WT}</i>	6.5	8	8
<i>Casp8^{C362S/C362S}</i>	3.25	3	1
Total	13	13	11



Extended Data Fig. 7 | ASC or caspase-1 deficiency rescues embryonic lethality of mice expressing CASP8(C362S). **a**, Expected, observed and weaned numbers of mice per genotype obtained from the indicated crossings. **b**, Representative images of 8-week-old mice (top) and spleen, axial and inguinal and mesenteric lymph nodes (middle) as well as splenic sections stained with H&E (bottom) from *Casp8^{C362S/WT} Mik1^{-/-} Asc^{-/-}* ($n = 3$), *Casp8^{C362S/C362S} Mik1^{-/-} Asc^{-/-}* ($n = 3$), *Casp8^{C362S/WT} Mik1^{-/-} Casp1^{-/-}* ($n = 3$) and *Casp8^{C362S/C362S} Mik1^{-/-} Casp1^{-/-}* ($n = 3$) 8-week-old mice. Scale bars, 300 μ m. **c**, Immunofluorescence

confocal images of endothelial cells treated with HTNCre and stained with an anti-ASC antibody after 24 h (top) Scale bar, 20 μ m. Measurement of IL-1 β levels in supernatants of endothelial cells after 24-h HTNCre treatment (bottom; $n = 3$ biologically independent replicates). Dots and circles represent individual biological replicates. Data are mean \pm s.e.m. One-way ANOVA followed by Sidak's post-analysis. Results are representative of two individual experiments. **d**, Ponceau staining of ileal lysates from 5-week-old mice ($n = 2$) (Fig. 4d). Lanes, individual mice.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Adobe Illustrator CS5 Version 15.1.0, Volocity Version 5.4.2 (PerkinElmer)

Data analysis GraphPad Prism 7.0 and Excel software were used for statistical analyses, LUNARIS™ Analysis Suite 1.3 (AYOXXA) for cytokine/chemokine analysis, Ape Plasmid Editor v2.0.49 Software for sequencing analysis, Aperio ImageScope Version v12.2.2.5015 (Leica) for scanned stained tissue section analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during the current study are available from the corresponding author on reasonable request

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For in vitro and in vivo experiments, sample sizes were chosen according to the basis of previous publications without prior power analysis. For mouse experiments we usually used 3 mice per group to ensure the statistically significant difference. On the other hand, we also tried to minimize the animal number.
Data exclusions	No data were excluded from the analysis in this study
Replication	Experimental findings were reliably reproduced and are representative of least two independent experiments
Randomization	We randomly chose the mice for each experiment group
Blinding	Embryos were imaged blindly before genotyping. Blinding was not possible as mice were sacrificed and evaluated by the same person.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

- 1) Monoclonal rabbit anti-caspase-3 (8G10), Cat. No. 9665, Cell Signaling, Dilution 1:1000, Lot. No. 1;
- 2) polyclonal rabbit anti-cleaved Caspase-3, Cat. No. 9661, Cell Signaling, Dilution 1:1000, Lot. No. 45;
- 3) polyclonal rabbit anti-caspase-8 (mouse specific), Cat. No. 4927, Cell Signaling, Dilution 1:1000, Lot. No. 2;
- 4) polyclonal rabbit anti-cleaved Caspase-8 (Asp387), Cat. No. 9429, Cell Signaling, Dilution 1:1000, Lot. No. 2;
- 5) monoclonal mouse anti-Caspase-8 (human specific) (1C12), Cat. No. 9746, Cell Signaling, Dilution 1:1000, Lot. No. 20;
- 6) polyclonal rabbit anti-Caspase-3 Cat. No. 9662, Cell Signaling, Dilution 1:1000, Lot. No. 18;
- 7) monoclonal mouse anti-Caspase-7 Cat. No. 9494, Cell Signaling, Dilution 1:1000, Lot. No. 2;
- 8) polyclonal rabbit anti-Caspase-9 Cat. No. 9502, Cell Signaling, Dilution 1:1000, Lot. No. 8;
- 9) monoclonal rat anti-Caspase-1 clone 5B10, Cat. No. 645102, Biolegend, Dilution 1:1000, Lot. No. B257128;
- 10) monoclonal mouse anti-RIPK1 clone G322-2, Cat. No. 51-6559GR, BD Biosciences, Dilution 1:1000, Lot. No. 87293;
- 11) polyclonal rabbit anti-RIPK3, Cat. No. ADI-905-242-100, Enzo, Dilution 1:1000, Lot. No. 04131703;
- 12) polyclonal rabbit anti-cFLIP alpha, Cat. No. F6550, Sigma Aldrich, Dilution 1:1000, Lot. No. 90K1043;
- 13) monoclonal mouse anti-FADD, Cat. No. F36620, BD Biosciences, Dilution 1:1000; Lot. No. 6
- 14) goat anti-rabbit IgG conjugated to horseradish peroxidase (HRP), Cat. No. 7074 purchased Cell Signaling, Dilution 1:1000, Lot. No. 28;
- 15) goat anti-mouse conjugated to HRP Cat. No. A4416, Sigma Aldrich, Dilution 1:3000;
- 16) goat anti-rabbit IgG biotin-labeled from Perkin Elmer, Cat. No. NEF813001EA, Lot. No. 10180764;
- 17) monoclonal rat anti-mouse CD31-Alexa Fluor 647 (MEC13.3) Cat. No. 102516 from BioLegend, Dilution 1:1000, Lot. No. B245809;
- 19) polyclonal rabbit anti-mouse Caspase-1 p10 (M-20), Cat. No. sc-514, Santa Cruz (discontinued), Dilution 1:500, Lot. No. D1415;
- 20) polyclonal rabbit anti-ASC (N15), Cat. No. sc-22514-R, Santa Cruz (discontinued), Dilution 1:1000 for WB and 1:200 for IHC;
- 21) monoclonal mouse anti-β-Actin HRP (C4) conjugated, Cat. No. sc-47778, Santa Cruz, Dilution 1:5000, Lot. No. K1418;
- 22) polyclonal rabbit anti-dsRed, Cat. No. 632397 was purchased from BD Biosciences (discontinued), Dilution 1:1000, Lot No.

403008;
 23) polyclonal rabbit anti-lysozyme EC.3.2.1.17, Cat. No. A0099 from DAKO, Dilution 1:2000, Lot. No. 20055029;
 24) polyclonal Goat anti-Rat IgG (H+L) secondary Antibody HRP, Cat. No 031470. from Invitrogen, Dilution 1:2000, Lot. No. UC280037;
 25) monoclonal rat anti-Caspase-8 (clone 1G12) (mouse-specific), Cat. No. ALX-804-447-C100 from Enzo, Dilution WB 1:1000 and IHC 1:200, Lot. No. 03111914;
 26) monoclonal mouse anti-Caspase-1 (p20) clone Casper-1, Cat. No. AG-20B-0042-C100 from Adipogen, Dilution 1:2000, Lot. No. A28881708;
 27) monoclonal rabbit anti-MLKL (phospho S345) EPR9515(2), Cat. No. ab196536, Abcam, Dilution 1:1000, Lot. No. GR3204546-2;
 28) goat anti-mouse IgG light chain HRP Cat. No. 115-035-174 from Jackson ImmunoResearch, Dilution 1:1000, Lot. No. 106128;
 29) goat anti-rabbit Alexa Fluor 568, Cat. No. A11036 from Thermo Fisher Scientific, Dilution 1:1000, Lot. No. 1832035
 30) goat anti-rabbit Alexa Fluor 594, Cat. No. A21442 from Thermo Fisher Scientific, Dilution 1:1000, Lot. No. 84E1-1
 31) Biotin-SP conj. AffiniPure Goat Anti-Rat IgG, Cat. No. 112065003 from Jackson ImmunoResearch, Dilution 1:1000, Lot. No. 139870

Validation

All antibodies were validated according to the manufacturers instruction and as described in the literature
 1) commercial antibody, Validation data available on manufacturers website, Pham, D. D., Bruelle, C., et al. 2019 Cell Death & Disease
 2) commercial antibody, Validation data available on manufacturers website, Schipper, K., Seinstra, D., et al. 2019 Nature Communications
 3) commercial antibody, Validation data available on manufacturers website, Martin, B. N., Wang, C., et al. 2016 Nature Immunology
 4) commercial antibody, Validation data available on manufacturers website, Taraborrelli, L., Peltzer, N., et al. 2018 Nature Communications
 5) commercial antibody, Validation data available on manufacturers website, Zhu, Z. C., Liu, J. W., et al. 2019 Cell Death & Disease
 6) commercial antibody, Validation data available on manufacturers website, Lazareth, H., Henique, C., et al. 2019 Nature Communications
 7) commercial antibody, Validation data available on manufacturers website, Breunig, C., Pahl, J., et al. 2017 Cell Death & Disease
 8) commercial antibody, Validation data available on manufacturers website, Füllsack, S., Rosenthal, A., et al. 2019 Cell Death & Disease
 9) Validated by Manufacturer, see Certificate of Analysis on website using Lot No B257128
 10) Validation statement in Datasheet of Product on Manufacturers website (routinely tested).
 11) Validated by Manufacturer, see Certificate of Analysis on website using Lot. No 04131703
 12) Validation statement on Manufacturers website (Spezifikation Sheet).
 13) commercial antibody, Validation data available on manufacturers website, Kischkel F.C. et. al. 2001, JBC
 17) Validated by Manufacturer, see Certificate of Analysis on website using Lot. No B245809
 19) commercial antibody, Validation data available on manufacturers website, Roth, S. et al. 2014. Nature immunology
 20) Antibody was validated for WB by ASC-/- MLKL-/- Casp8C362S/C362S ileal lysates in Figure 4d and for IHC in ASC-/- MLKL-/- Casp8C362S/C362S ileal sections in Figure 4b
 21) commercial antibody, Validation data available on manufacturers website, Shan, Y. et al. 2019, Cell Death Disease
 22) Antibody was validated for WB by DsRed overexpression vector in cellular lysates in Figure 3d
 23) commercial antibody, Validation data available on manufacturers website, Huels, D. J., Bruens, L., et al. 2018 Nature Communications
 25) Validated by Manufacturer, see Certificate of Analysis on website using Lot. No 03111914
 26) commercial antibody, Validation data available on manufacturers website, Karki R. et al. 2018 Cell
 27) Validation statement on Manufacturers website for WB

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Isolated primary endothelial cells by magnetic bead separation (murine CD31), isolated bone-marrow derived macrophages (BMDMs), HCT-116 and HEK293T cell lines were purchased from ATCC

Authentication

Isolated primary endothelial cells were analysed using CD31 antibody, sprouting response and mRNA levels to distinguish endothelial cells from other cell types; BMDM were isolated and differentiated according to the current literature; Casp8 KO cell lines were determined via WB analysis as well as sequencing. HCT-116 and HEK293T cell lines were purchased from ATCC.

Mycoplasma contamination

All cell lines were tested routinely negative for Mycoplasma contamination by PCR

Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified lines were used.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

Mus musculus C57BL/6N mixed sexes, 1-30 weeks old according to the figure legends

Wild animals

This study did not involve wild animals

Field-collected samples

This study did not involve field-collected samples

Ethics oversight

All mouse studies were performed after approval by local government authorities (LANUV, NRW, Germany) in accordance with the German animal protection law.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

The CoQ oxidoreductase FSP1 acts parallel to GPX4 to inhibit ferroptosis

<https://doi.org/10.1038/s41586-019-1705-2>

Received: 25 February 2019

Accepted: 1 October 2019

Published online: 21 October 2019

Kirill Bersuker¹, Joseph M. Hendricks¹, Zhipeng Li¹, Leslie Magtanong², Breanna Ford^{1,3,4}, Peter H. Tang⁵, Melissa A. Roberts¹, Bingqi Tong³, Thomas J. Maimone³, Roberto Zoncu⁴, Michael C. Bassik⁶, Daniel K. Nomura^{1,3,4}, Scott J. Dixon² & James A. Olzmann^{1,7*}

Ferroptosis is a form of regulated cell death that is caused by the iron-dependent peroxidation of lipids^{1,2}. The glutathione-dependent lipid hydroperoxidase glutathione peroxidase 4 (GPX4) prevents ferroptosis by converting lipid hydroperoxides into non-toxic lipid alcohols^{3,4}. Ferroptosis has previously been implicated in the cell death that underlies several degenerative conditions², and induction of ferroptosis by the inhibition of GPX4 has emerged as a therapeutic strategy to trigger cancer cell death⁵. However, sensitivity to GPX4 inhibitors varies greatly across cancer cell lines⁶, which suggests that additional factors govern resistance to ferroptosis. Here, using a synthetic lethal CRISPR–Cas9 screen, we identify ferroptosis suppressor protein 1 (FSP1) (previously known as apoptosis-inducing factor mitochondrial 2 (AIFM2)) as a potent ferroptosis-resistance factor. Our data indicate that myristoylation recruits FSP1 to the plasma membrane where it functions as an oxidoreductase that reduces coenzyme Q₁₀ (CoQ) (also known as ubiquinone-10), which acts as a lipophilic radical-trapping antioxidant that halts the propagation of lipid peroxides. We further find that FSP1 expression positively correlates with ferroptosis resistance across hundreds of cancer cell lines, and that FSP1 mediates resistance to ferroptosis in lung cancer cells in culture and in mouse tumour xenografts. Thus, our data identify FSP1 as a key component of a non-mitochondrial CoQ antioxidant system that acts in parallel to the canonical glutathione-based GPX4 pathway. These findings define a ferroptosis suppression pathway and indicate that pharmacological inhibition of FSP1 may provide an effective strategy to sensitize cancer cells to ferroptosis-inducing chemotherapeutic agents.

GPX4 is considered to be the primary enzyme that prevents ferroptosis². The resistance of some cancer cell lines to GPX4 inhibitors⁶ led us to search for additional protective pathways. To identify ferroptosis-resistance genes, we performed a synthetic lethal CRISPR–Cas9 screen using a sublibrary of single-guide RNAs (sgRNAs) targeting genes related to apoptosis and cancer in U-2 OS osteosarcoma cells that were treated with the GPX4 inhibitor 1S,3R-RSL3 (hereafter, RSL3) (Fig. 1a). This screen revealed a substantial dis-enrichment of sgRNAs targeting *FSP1* (currently known as *AIFM2*) in the cells treated with RSL3 (Fig. 1b, c, Extended Data Fig. 1a, Supplementary Table 1), which indicates that deletion of the *FSP1* gene is lethal in combination with RSL3 treatment. FSP1 was originally named AIFM2 on the basis of its homology with apoptosis-inducing factor (AIF or AIFM1), a mitochondrial pro-apoptotic protein^{7,8}. However, as we report here, FSP1 lacks the N-terminal mitochondrial targeting sequence in AIF, does not localize to mitochondria and does not promote apoptosis. We rename AIFM2 as FSP1 to reflect its cellular role, as described in this study.

FSP1 is a potent ferroptosis suppressor

Quantification of cell viability using time-lapse microscopy revealed a considerable increase in the sensitivity of FSP1 knockout (FSP1^{KO}) cell lines to RSL3 (Fig. 1d–f, Extended Data Fig. 1b, Supplementary Table 2), which was rescued by expression of untagged FSP1 (Extended Data Fig. 1c, d). In contrast to previous reports^{7,8}, the overexpression of FSP1 did not induce apoptosis (Extended Data Fig. 1e, f) and activation of p53 did not increase FSP1 expression (Extended Data Fig. 1g). FSP1^{KO} cells displayed increased sensitivity to additional ferroptosis inducers, including the GPX4 inhibitor ML162 and the system x_c[−] inhibitor erastin2 (ref.9) (Extended Data Fig. 1h), but not to the complex I inhibitor rotenone or hydrogen peroxide (Extended Data Fig. 1i–l). The viability of RSL3-treated FSP1^{KO} cells was rescued by the iron chelator deferoxamine (DFO) and by the radical-trapping antioxidants ferrostatin-1 (Fer1) and idebenone (Fig. 1g), but not by inhibitors of apoptosis (ZVAD(OMe)-FMK) or necroptosis (necrostatin-1) (Extended Data Fig. 1m). Knockout of long-chain acyl-CoA synthetase 4 (ACSL4) in

¹Department of Nutritional Sciences and Toxicology, University of California, Berkeley, Berkeley, CA, USA. ²Department of Biology, Stanford University, Stanford, CA, USA. ³Department of Chemistry, University of California, Berkeley, Berkeley, CA, USA. ⁴Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. ⁵Department of Pathology and Laboratory Medicine, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH, USA. ⁶Department of Genetics and Stanford University Chemistry, Engineering and Medicine for Human Health (ChEM-H), Stanford University School of Medicine, Stanford, CA, USA. ⁷Chan Zuckerberg Biohub, San Francisco, CA, USA.

*e-mail: olzmann@berkeley.edu

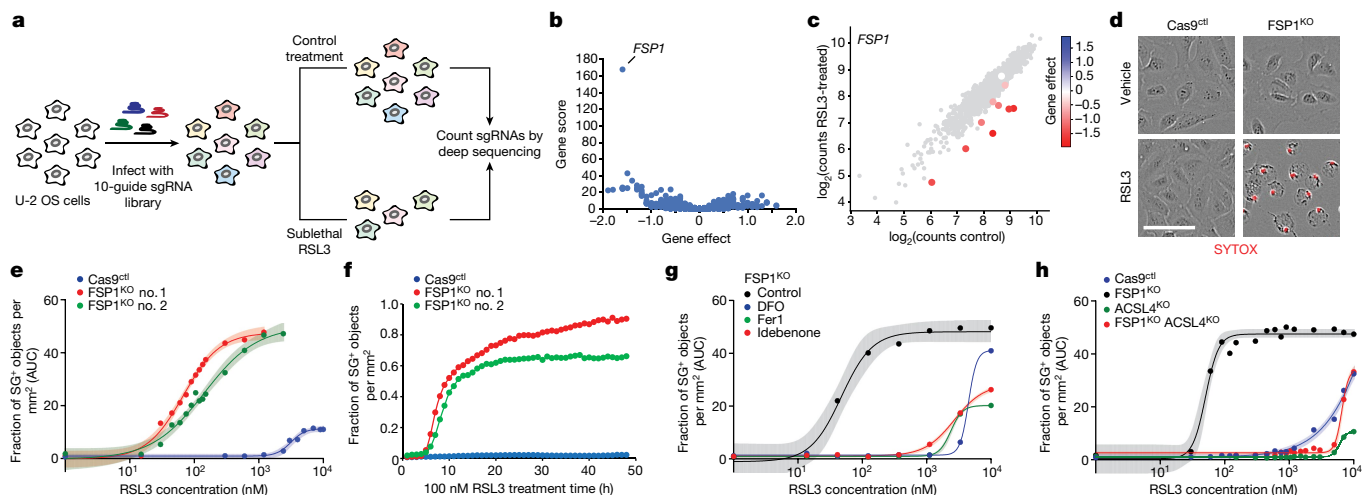


Fig. 1 | A synthetic lethal CRISPR–Cas9 screen identifies FSP1 as a ferroptosis resistance factor. **a**, Schematic of the CRISPR–Cas9 screening strategy. **b**, Gene effect and gene score calculated for individual genes analysed in the CRISPR–Cas9 screen. **c**, Cloud plot indicating count numbers corresponding to *FSP1* (colour scale) and control (grey) sgRNAs. The gene effect of individual sgRNAs targeting *FSP1* is indicated by the heat map. **d**, Live-cell imaging of control (Cas9^{ctrl}) and FSP1^{KO} cells incubated with SYTOX Green (SG*) and treated with 100 nM RSL3 for 48 h. Scale bar, 75 μ m. **e**, Dose response of RSL3-induced cell death of control and FSP1^{KO} cells. AUC, area under the curve. **f**, Time-lapse

cell death analysis of cells treated with 100 nM RSL3 over 48 h. **g**, Dose response of RSL3-induced cell death in the presence of inhibitors of ferroptosis (Fer1, 1 μ M; DFO, 100 μ M; and idebenone, 10 μ M). **h**, Dose response analysis of RSL3-induced cell death of the indicated cell lines. The ACSL4^{KO} and ACSL4^{KO} FSP1^{KO} lines shown were generated using ACSL4 sgRNA no. 1. In **e**, **g**, **h**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. Panels are representative of two biological replicates, except for **b** and **c**, which were derived from a single screen.

FSP1^{KO} cells (FSP1^{KO} ACSL4^{KO}) restored resistance to RSL3 to an extent similar to that of knockout of ACSL4 alone (ACSL4^{KO}) (Fig. 1h, Extended Data Fig. 1n), consistent with the requirement for ACSL4-mediated incorporation of polyunsaturated fatty acids into phospholipids for ferroptosis^{2,5}. Together, these findings demonstrate that FSP1 is a strong suppressor of ferroptosis.

Plasma-membrane FSP1 blocks ferroptosis

FSP1 contains a short N-terminal hydrophobic sequence and a canonical flavin adenine dinucleotide-dependent oxidoreductase domain (Extended Data Fig. 1o). FSP1 has previously been detected on lipid droplets¹⁰. To further define the localization of FSP1, we inserted a C-terminal halotag (HaloTag) into the *FSP1* genomic locus (Fig. 2a). Similar to ectopically expressed wild-type FSP1 tagged at the C terminus with green fluorescent protein (FSP1(WT)–GFP) (Extended Data Fig. 2a,b), FSP1–HaloTag localized to the periphery of lipid droplets and to the plasma membrane (Fig. 2b, Extended Data Fig. 2c, d). FSP1–HaloTag did not co-localize with endoplasmic reticulum labelled with blue fluorescent protein fused to Sec61 (BFP–Sec61) or with mitochondria labelled with MitoTracker (Extended Data Fig. 2g, h), consistent with the absence of an endoplasmic reticulum or mitochondrial targeting motif in FSP1. We noted an N-terminal consensus sequence for myristoylation (Fig. 2c), a fatty acid modification that is known to function in membrane targeting. FSP1 myristoylation was tested using a click chemistry method that enables affinity purification of myristoylated proteins (Extended Data Fig. 3a). Using this approach, endogenous FSP1 was affinity-purified from buoyant fractions enriched in lipid droplets (Extended Data Fig. 3b) and from whole-cell lysates (Fig. 2d). The myristoylation of FSP1(WT)–GFP was blocked by the inhibition of *N*-myristoyltransferase (NMT), mutation of the glycine-2 of FSP1 to alanine (FSP1(G2A)–GFP) and treatment with the translation inhibitor emetine (Fig. 2d, Extended Data Fig. 3c). Chemical and genetic perturbations of FSP1 myristoylation blocked FSP1 recruitment to lipid droplets (Fig. 2e, Extended Data Fig. 3d, e). Although a portion of FSP1(G2A)–GFP was observed in proximity to the plasma membrane by total internal reflection fluorescence (TIRF) microscopy

(Extended Data Fig. 2f), the fractionation of organelles in iodixanol (OptiPrep) gradients revealed that FSP1(G2A)–GFP was present at lower levels in fractions enriched in plasma membrane (Fig. 2f, g, Extended Data Fig. 2i). Together, these results indicate that the myristoylation of FSP1 mediates the recruitment of this protein to lipid droplets and the plasma membrane.

Expression of FSP1(WT)–GFP, but not of FSP1(G2A)–GFP, rescued the resistance of FSP1^{KO} cells to RSL3 (Fig. 2h, Extended Data Fig. 3f), which indicates that FSP1 must be myristoylated to suppress ferroptosis. We generated fusion proteins that selectively target FSP1(G2A)–GFP to the endoplasmic reticulum (amino acids 100–134 of cytochrome *b5*; Cb5), the outer mitochondrial membrane (TOM20 signal sequence, TOM20(SS)), lipid droplets (PLIN2) and the plasma membrane (first 11 amino acids of LYN kinase; LYN11) (Extended Data Fig. 4a, b). Only the expression of FSP1 targeted to the plasma membrane (LYN11–FSP1(G2A)–GFP) was sufficient to restore ferroptosis resistance in FSP1^{KO} cells (Fig. 2i, Extended Data Fig. 4c). By contrast, expression of FSP1(G2A)–GFP targeted to the endoplasmic reticulum, mitochondria or lipid droplets had no effect (Fig. 2i). Consistent with previous results in HT1080 cells¹¹, the depletion of lipid droplets using inhibitors of the diacylglycerol acyltransferase enzymes (DGAT1 and DGAT2) did not affect ferroptosis sensitivity (Extended Data Fig. 5a–c), which provides support for the conclusion that lipid-droplet localization is not required for the FSP1-mediated suppression of ferroptosis. Thus, FSP1 plasma-membrane localization is necessary and sufficient to confer ferroptosis resistance.

FSP1 reduces CoQ to suppress ferroptosis

Under basal conditions, the ratiometric fluorescent lipid peroxidation sensor BODIPY 581/591 C11 exhibited similar levels of oxidation in control and FSP1^{KO} cells (Fig. 3a, Extended Data Fig. 6a, b). However, a brief treatment with RSL3 strongly increased C11 oxidation in FSP1^{KO} cells relative to control (Fig. 3a, Extended Data Fig. 6a, b). Glutathione levels were unaffected in FSP1^{KO} cells (Extended Data Fig. 6c–e), indicating that deletion of FSP1 does not inhibit system *x*_c[−] or glutathione synthesis. FSP1^{KO} cells also did not exhibit higher levels of phospholipids

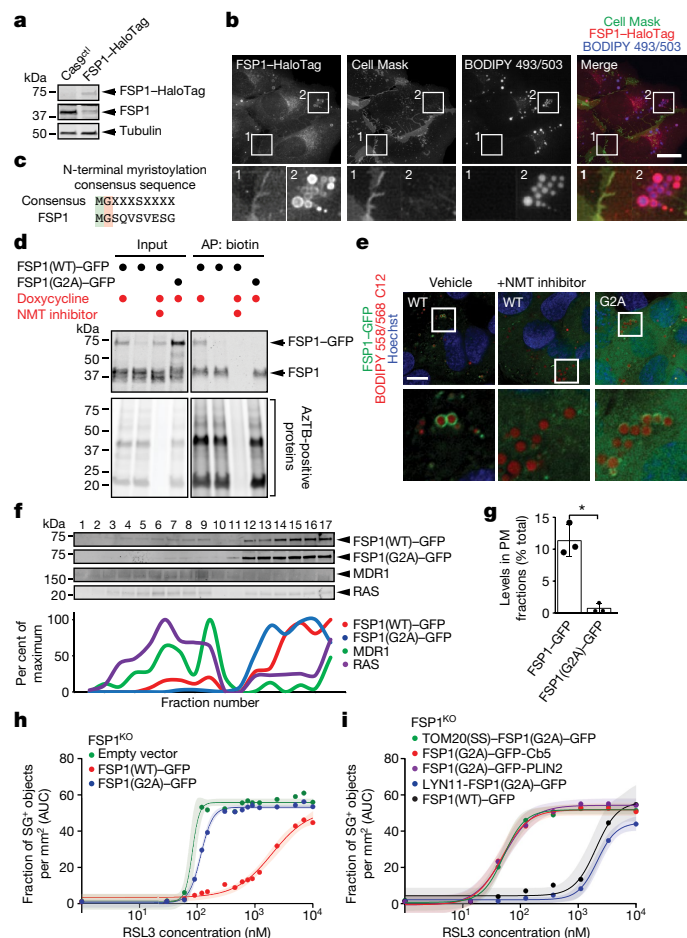


Fig. 2 | Myristoylation-dependent recruitment of FSP1 to the plasma membrane promotes ferroptosis resistance. a, Western blot of lysates from FSP1-HaloTag genomic knock-in cells. **b**, FSP1-HaloTag subcellular distribution by live-cell microscopy. Cells were incubated with 100 nM JF549 to label FSP1-HaloTag, 5 $\mu\text{g ml}^{-1}$ Cell Mask to label the plasma membrane and 1 $\mu\text{g ml}^{-1}$ BODIPY 493/503 to label lipid droplets. **c**, Consensus myristoylation sequence in FSP1. **d**, Analysis of FSP1-GFP myristoylation in whole-cell lysates of the indicated cell lines treated for 24 h with doxycycline to induce FSP1-GFP expression. Where indicated, 10 μM NMT inhibitor was added for 24 h to inhibit myristoylation. AP, affinity purification; AzTB, TAMRA-azide-PEG-biotin. **e**, Live-cell microscopy of inducible FSP1-GFP cell lines treated with 200 μM oleate and 1 μM BODIPY 558/568 C12. Where indicated, cells were treated concurrently with 10 μM NMT inhibitor. **f**, Subcellular fractionation of organelles from cells that express FSP1-GFP, using OptiPrep gradient centrifugation. The densitometry plot shows the distribution of the indicated overexpressed and endogenous proteins. **g**, Quantification of FSP1-GFP levels in fractions 1–10 in **f**. The graph shows mean \pm s.d. of $n = 3$ biological replicates. $^*P = 0.0124$ by two-tailed t -test. PM, plasma membrane. **h**, Dose response of RSL3-induced death of FSP1^{KO} cells pretreated with doxycycline for 48 h to induce expression of the indicated FSP1-GFP proteins. **i**, Dose response of RSL3-induced death of FSP1^{KO} cells that express the indicated inducible FSP1(G2A)-GFP constructs. In **h**, **i**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. All panels are representative of two biological replicates. Images are representative of at least $n = 10$ imaged cells. Scale bars, 10 μm .

that contain polyunsaturated fatty acids (Extended Data Fig. 6f, g, Supplementary Table 3). Levels of phospholipids containing polyunsaturated fatty acids were decreased and the corresponding lysophospholipids were increased (Extended Data Fig. 6f, g, Supplementary Table 3), a known lipidomic signature of ferroptosis that reflects the removal of oxidized polyunsaturated fatty acids from the *sn*-2 position

of phospholipids^{3,12}. These results suggest that the loss of FSP1 results in increased phospholipid oxidation even when GPX4 is functional, and that FSP1 prevents lipid peroxidation through a mechanism that is distinct from glutathione-dependent protective pathways.

FSP1 functions as an NADH-dependent CoQ oxidoreductase *in vitro*¹³. Reduced CoQ can act as a radical-trapping antioxidant, and idebenone—a soluble analogue of CoQ—is sufficient to suppress lipid peroxidation (Extended Data Fig. 7a) and ferroptosis¹⁴ (Fig. 1g). Previous studies have detected high levels of CoQ in non-mitochondrial compartments, including the plasma membrane^{15,16}, but the function of this molecule in these compartments remains unclear. To examine the role of FSP1 CoQ oxidoreductase activity in suppressing ferroptosis, we mutated a conserved glutamate residue (E13 in AIF or E156 in FSP1) that is required for the binding of AIF to its cofactor, flavin adenine dinucleotide (Extended Data Fig. 7b, c). Mutation of E156 in FSP1 (FSP1(E156A)-GFP) did not affect FSP1-GFP expression or localization (Extended Data Figs. 3f, 7d, e) but greatly impaired FSP1-mediated reduction of coenzyme Q₁ and resazurin *in vitro* (Extended Data Fig. 7f–h) and abolished the ability of FSP1-GFP to rescue the resistance of FSP1^{KO} cells to RSL3 (Fig. 3b). Consistent with these findings, the expression of FSP1(WT)-GFP, but not of FSP1(E156A)-GFP, increased the ratio of reduced-to-oxidized CoQ (Fig. 3c). Acute reduction of cellular CoQ levels by the inhibition of the CoQ biosynthesis enzyme COQ2 with 4-chlorobenzoic acid (4-CBA) strongly sensitized control cells and—to a lesser extent—FSP1^{KO} cells to RSL3-induced ferroptosis (Fig. 3d, e, Extended Data Fig. 8a). Treatment with 4-CBA also suppressed the ability of FSP1(WT)-GFP to rescue FSP1^{KO} cells (Extended Data Fig. 8b). A similar degree of sensitization to RSL3 was observed after knockout of COQ2 in control, but not in FSP1^{KO}, cells (Fig. 3f, g, Extended Data Fig. 8c) and COQ2^{KO} cells exhibited increased C11 oxidation after treatment with RSL3 that was suppressed by DFO and by idebenone (Extended Data Fig. 8d, e). These data indicate that FSP1 and the CoQ synthesis machinery function in the same pathway to suppress lipid peroxidation and ferroptosis.

Deletion of NQO1, a quinone and CoQ oxidoreductase that has previously been proposed to function in ferroptosis¹⁷, did not affect sensitivity to RSL3, but cells that lack both FSP1 and NQO1 (FSP1^{KO} NQO1^{KO}) were more sensitive than FSP1^{KO} cells (Extended Data Fig. 9a–c). NQO1-GFP did not rescue ferroptosis resistance in FSP1^{KO} cells to the same extent as did FSP1-GFP (Extended Data Fig. 9d–g), even when targeted to the plasma membrane (LYN11-NQO1-GFP) (Extended Data Fig. 9h, i). These results indicate that FSP1 is unique in its ability to suppress ferroptosis through the reduction of CoQ.

FSP1 in cancer ferroptosis resistance

The Cancer Therapeutics Response Portal (CTRP) reports correlations between gene expression and drug resistance for over 800 cancer cell lines¹⁸. Data mined from the CTRP indicate that FSP1 expression positively correlates with resistance to multiple GPX4 inhibitors (RSL3, ML210 and ML162) (Fig. 4a, b, Extended Data Fig. 10a, b, Supplementary Table 4)—even more so than the system x_c^- component and erastin target SLC7A11⁹. Thus, FSP1 is a biomarker of ferroptosis resistance in many types of cancer. Consistent with the correlations observed in the CTRP, lung cancer cell lines that express low levels of FSP1 were the most sensitive to RSL3 and cell lines that express high levels of FSP1 were the most resistant (Fig. 4b, Extended Data Fig. 10c). Knock-out of FSP1 in the highly resistant H460 cell line resulted in a notable, approximately 100-fold sensitization to RSL3 (Fig. 4d, Extended Data Fig. 10d, e) and overexpression of FSP1-GFP in sensitive H1703 and H446 cells increased resistance to RSL3 by about 10–20 fold (Fig. 4e, Extended Data Fig. 10f–i).

To examine the possibility that the inhibition of FSP1 could be a clinically relevant approach to sensitize tumours to ferroptosis-activating chemotherapies, we used ferroptosis-resistant H460 lung cancer cells in a preclinical tumour xenograft mouse model. Owing to the poor

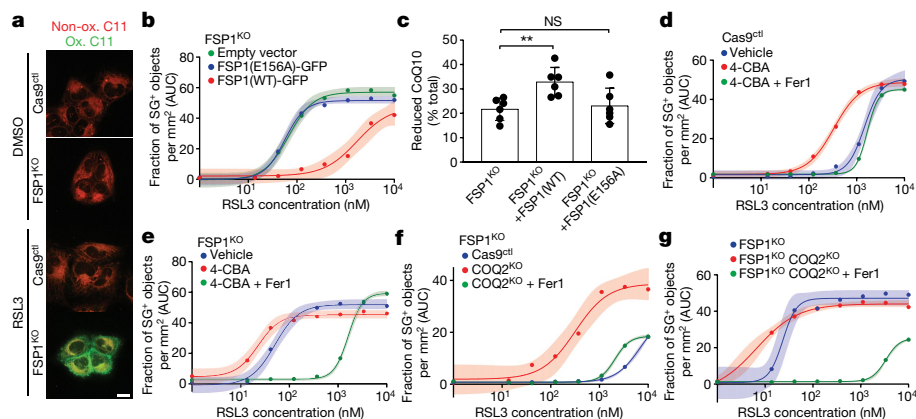


Fig. 3 | FSP1 suppresses lipid peroxidation by reducing CoQ. **a**, Control and FSP1^{KO} cells treated with 250 nM RSL3 for 75 min were labelled with BODIPY 581/591 C11 and fixed before imaging. Ox., oxidized; Non-ox., non-oxidized. Images are representative of at least 30 cells imaged for each treatment condition. Scale bar, 20 μ m. **b**, Dose response of RSL3-induced cell death of FSP1^{KO} cells that express the indicated inducible FSP1–GFP constructs. **c**, Reduced-to-oxidized CoQ ratio in FSP1^{KO} and FSP1^{KO} cells that express the indicated FSP1–GFP constructs. Data represent mean \pm s.d. of $n = 6$ biological

replicates. ** $P = 0.0178$, NS, not significant ($P > 0.99$), by one-way analysis of variance (ANOVA). **d**, **e**, Dose response of RSL3-induced death of control (**d**) and FSP1^{KO} (**e**) cells pretreated for 24 h with 3 mM 4-CBA. **f**, **g**, Dose response of RSL3-induced cell death of COQ2^{KO} (**f**) and FSP1^{KO} COQ2^{KO} (**g**) cells. In **b**, **d**–**g**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. All figures are representative of two biological replicates.

bioavailability of small-molecule GPX4 inhibitors (such as RSL3), we adopted a recently developed strategy to acutely induce ferroptosis in vivo^{19,20} using GPX4 knockout (GPX4^{KO}) cells (Extended Data Fig. 10j). These cell lines were maintained in Fer1-containing medium to prevent the induction of ferroptosis. Fer1 washout had no effect on the viability of the GPX4^{KO} cells but resulted in the rapid death of GPX4^{KO} FSP1^{KO}

cells (Fig. 4f), consistent with our findings that FSP1 compensates for loss of GPX4 activity. Tumour xenografts were generated with GPX4^{KO} and GPX4^{KO} FSP1^{KO} H460 cell lines, and Fer1 was injected daily to allow viable tumours to develop. After tumours were established, Fer1 injections were discontinued in one set of mice to induce ferroptosis. In contrast to the GPX4^{KO} tumours (which continued to increase in size

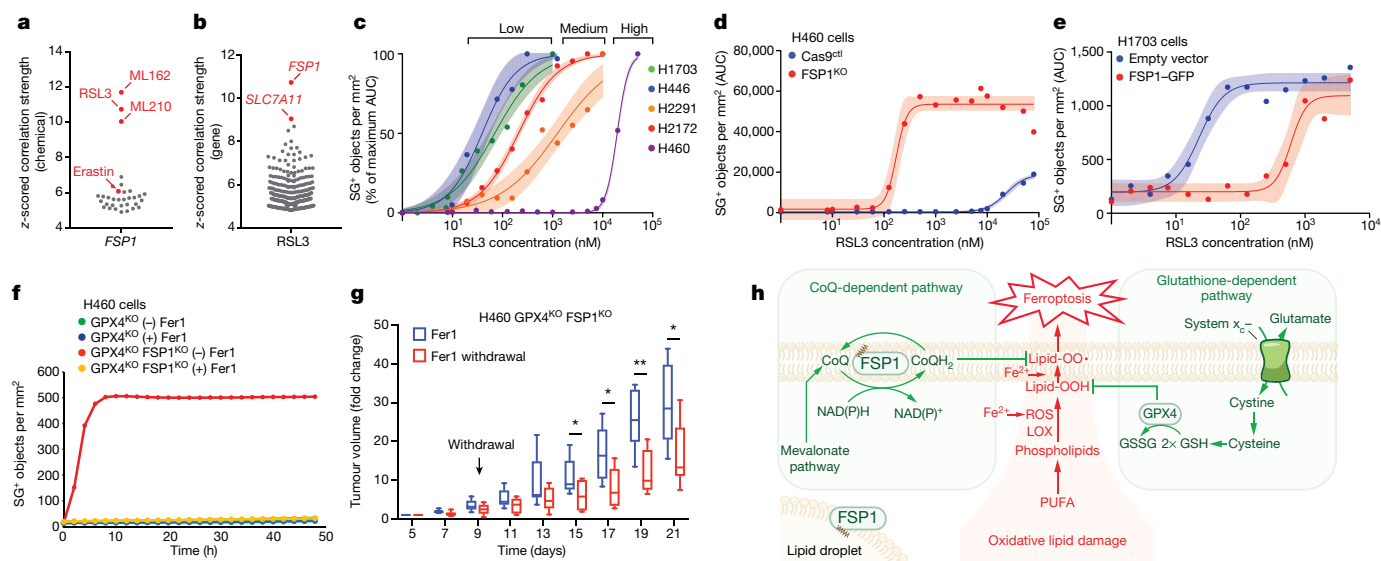


Fig. 4 | FSP1 mediates ferroptosis resistance in lung cancer. **a**, **b**, A high level of expression of FSP1 is correlated with resistance to GPX4 inhibitors in non-haematopoietic cancer cells. Plotted data were mined from the CTRP database, which contains correlation coefficients between gene expression and drug sensitivity for 907 cancer cell lines treated with 545 compounds. **a**, Correlation between FSP1 expression and resistance to individual compounds. **b**, Correlation between expression levels of individual genes and resistance to RSL3. Plotted values are z-scored Pearson's correlation coefficients. **c**, Dose response of RSL3-induced cell death of the indicated cell lines. **d**, Dose response of RSL3-induced cell death of control and FSP1^{KO} H460 cells. **e**, Dose response of RSL3-induced cell death of FSP1–GFP H1703 cells. **f**, Time-lapse analysis of cell death of GPX4^{KO} and GPX4^{KO} FSP1^{KO} H460 cells in the presence and absence of 1 μ M Fer1. **g**, GPX4^{KO} FSP1^{KO} H460 tumour xenograft cells were

initiated in immune-deficient SCID mice ($n = 16$). Following 5 days of daily Fer1 injections (2 mg kg⁻¹ body weight) to allow the cell lines to develop tumours, one set of mice ($n = 8$) continued to receive daily Fer1 injections and a second set ($n = 8$) received vehicle injections for the remaining 17 days. The distribution of fold changes in sizes of individual tumours during the treatment is shown. GPX4^{KO} FSP1^{KO} (–) Fer1, $n = 7$; GPX4^{KO} FSP1^{KO} (+) Fer1, $n = 8$. Box plots indicate median, 25th and 75th percentiles, and minima and maxima of the distributions. Day 15, * $P = 0.0397$; day 17, * $P = 0.0187$; day 18, ** $P = 0.0025$; day 21, * $P = 0.0327$ by two-tailed t -test. **h**, Model illustrating the mechanism by which FSP1 suppresses ferroptosis. In **c**–**e**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. Panels **c**–**f** are representative of two biological replicates.

irrespective of Fer1; Extended Data Fig. 10k), Fer1 withdrawal resulted in a significant reduction in the growth of the GPX4^{KO} FSP1^{KO} tumours (Fig. 4g). These data demonstrate that FSP1 maintains the growth of H460 lung cancer tumours in vivo when GPX4 is inactivated. To determine whether the growth of FSP1^{KO} tumours can be inhibited by blocking cystine import, we treated H460 cells with imidazole-ketone-erastin (IKE), a system x_c⁻ inhibitor that can induce ferroptosis in vivo²¹. Although U-2 OS and H460 FSP1^{KO} cells exhibited increased sensitivity to IKE in cell culture (Extended Data Fig. 10i, m), IKE did not inhibit the growth of wild-type H460 and H460 FSP1^{KO} tumour xenografts (Extended Data Fig. 10n, o). Because cells can overcome the effects of cystine depletion through the use of alternative pathways to generate glutathione²², our results underscore the need for GPX4 inhibitors that are efficacious in vivo.

Ferroptosis has emerged as a potential cause of cell death in degenerative diseases and as a promising strategy to induce the death of cancer cells that are resistant to other therapies^{1,2,5}. Our studies and those of a companion paper²³ identify FSP1 as a potent ferroptosis suppressor that operates in parallel to the canonical glutathione-dependent GPX4 pathway. FSP1^{KO} mice are viable and display no obvious mutant phenotypes²⁴, consistent with the compensatory suppression of lipid peroxidation by GPX4. Mechanistically, our data support a model in which myristoylation targets FSP1 to the plasma membrane where it mediates the NADH-dependent reduction of CoQ, which functions as a radical-trapping antioxidant that suppresses the propagation of lipid peroxides (Fig. 4h). Our data also reveal that a fundamental role of non-mitochondrial CoQ is to function as an antioxidant that prevents lipid damage, and consequently ferroptosis. Localization of FSP1 at lipid droplets is not required for protection from ferroptosis. One possibility is that the FSP1-mediated regulation of lipophilic radical-trapping antioxidants in lipid droplets is important for the maintenance of lipid quality during prolonged periods of lipid storage, similar to the function of CoQ and tocopherol in preventing the oxidation of circulating lipoprotein particles^{25,26}. Finally, our findings indicate that FSP1 expression is important for predicting the efficacy of ferroptosis-inducing drugs in cancers and highlight the potential for FSP1 inhibitors²³ as a strategy to overcome ferroptosis resistance in multiple types of cancer.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1705-2>.

- Dixon, S. J. et al. Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* **149**, 1060–1072 (2012).
- Stockwell, B. R. et al. Ferroptosis: a regulated cell death nexus linking metabolism, redox biology, and disease. *Cell* **171**, 273–285 (2017).
- Yang, W. S. et al. Regulation of ferroptotic cancer cell death by GPX4. *Cell* **156**, 317–331 (2014).
- Ingold, I. et al. Selenium utilization by GPX4 is required to prevent hydroperoxide-induced ferroptosis. *Cell* **172**, 409–422.e21 (2018).
- Dixon, S. J. & Stockwell, B. R. The hallmarks of ferroptosis. *Annu. Rev. Cancer Biol.* **3**, 35–54 (2019).
- Zou, Y. et al. A GPX4-dependent cancer cell state underlies the clear-cell morphology and confers sensitivity to ferroptosis. *Nat. Commun.* **10**, 1617 (2019).
- Wu, M., Xu, L.-G., Li, X., Zhai, Z. & Shu, H.-B. AMID, an apoptosis-inducing factor-homologous mitochondrion-associated protein, induces caspase-independent apoptosis. *J. Biol. Chem.* **277**, 25617–25623 (2002).
- Ohno, Y. et al. A novel p53-inducible apoptogenic gene, *PRG3*, encodes a homologue of the apoptosis-inducing factor (AIF). *FEBS Lett.* **524**, 163–171 (2002).
- Dixon, S. J. et al. Pharmacological inhibition of cysteine–glutamate exchange induces endoplasmic reticulum stress and ferroptosis. *eLife* **3**, e02523 (2014).
- Bersuker, K. et al. A Proximity labeling strategy provides insights into the composition and dynamics of lipid droplet proteomes. *Dev. Cell* **44**, 97–112.e7 (2018).
- Magtanong, L. et al. Exogenous monounsaturated fatty acids promote a ferroptosis-resistant cell state. *Cell Chem. Biol.* **26**, 420–432.e9 (2019).
- Yang, W. S. et al. Peroxidation of polyunsaturated fatty acids by lipoxygenases drives ferroptosis. *Proc. Natl Acad. Sci. USA* **113**, E4966–E4975 (2016).
- Marshall, K. R. et al. The human apoptosis-inducing protein AMID is an oxidoreductase with a modified flavin cofactor and DNA binding activity. *J. Biol. Chem.* **280**, 30735–30740 (2005).
- Shimada, K. et al. Global survey of cell death mechanisms reveals metabolic regulation of ferroptosis. *Nat. Chem. Biol.* **12**, 497–503 (2016).
- Arroyo, A., Navarro, F., Navas, P. & Villalba, J. M. Ubiquinol regeneration by plasma membrane ubiquinone reductase. *Protoplasma* **205**, 107–113 (1998).
- Takahashi, T., Okamoto, T., Mori, K., Sayo, H. & Kishi, T. Distribution of ubiquinone and ubiquinol homologues in rat tissues and subcellular fractions. *Lipids* **28**, 803–809 (1993).
- Sun, X. et al. Activation of the p62-Keap1-NRF2 pathway protects against ferroptosis in hepatocellular carcinoma cells. *Hepatology* **63**, 173–184 (2016).
- Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
- Hangauer, M. J. et al. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature* **551**, 247–250 (2017).
- Viswanathan, V. S. et al. Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* **547**, 453–457 (2017).
- Zhang, Y. et al. Imidazole ketone erastin induces ferroptosis and slows tumor growth in a mouse lymphoma model. *Cell Chem. Biol.* **26**, 623–633.e9 (2019).
- Hayano, M., Yang, W. S., Corn, C. K., Pagano, N. C. & Stockwell, B. R. Loss of cysteinyl-tRNA synthetase (CARS) induces the transsulfuration pathway and inhibits ferroptosis induced by cystine deprivation. *Cell Death Differ.* **23**, 270–278 (2016).
- Doll, S. et al. FSP1 is a glutathione-independent ferroptosis suppressor. *Nature* <https://doi.org/10.1038/s41586-019-1707-0> (2019).
- Nguyen, T. B. et al. DGAT1-dependent lipid droplet biogenesis protects mitochondrial function during starvation-induced autophagy. *Dev. Cell* **42**, 9–21.e5 (2017).
- Tribble, D. L. et al. Oxidative susceptibility of low density lipoprotein subfractions is related to their ubiquinol-10 and α -tocopherol content. *Proc. Natl Acad. Sci. USA* **91**, 1183–1187 (1994).
- Stocker, R., Bowry, V. W. & Frei, B. Ubiquinol-10 protects human low density lipoprotein more efficiently against lipid peroxidation than does α -tocopherol. *Proc. Natl Acad. Sci. USA* **88**, 1646–1650 (1991).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines and culture conditions

U-2 OS T-Rex Flp-In cells, a gift from D. Durocher, and U-2 OS Tet-On cells (Clontech) were cultured in DMEM containing 4.5 g/l glucose and L-glutamine (Corning). NCI-H460, NCI-H2291, NCI-H1703 and NCI-H446 cells (ATCC) were cultured in RPMI1640 containing high glucose, L-glutamine and HEPES (ATCC). U-2 OS COQ2^{ko} cells were grown in DMEM supplemented with 200 μ M uridine and FSP1^{ko} COQ2^{ko} cells were grown in DMEM supplemented with 200 μ M uridine and 1 μ g/ml Fer1. NCI-H460 GPX4^{ko} lines and FSP1^{ko} GPX4^{ko} lines were grown in RPMI1640 supplemented with 1 μ g/ml Fer1. All media were supplemented with 10% fetal bovine serum (FBS, Thermo Fisher Scientific and Gemini Bio Products), and all cell lines were grown at 37 °C with 5% CO₂. All cell lines were tested for mycoplasma and were not authenticated.

Generation of doxycycline-inducible cell lines

U-2 OS expression lines were generated by transfection of U-2 OS T-Rex Flp-In cells with pOG44 Flp-Recombinase plasmid (Thermo Fisher Scientific) and pcDNA5/FRT/TO plasmid at a 9:1 ratio, followed by selection in 500 μ g/ml hygromycin. NCI-H1703 and NCI-H446 expression lines were generated by infection with pLenti CMV TetR Blast virus (716-1) (Addgene plasmid no. 17492) in the presence of 8 μ g/ml polybrene (Sigma-Aldrich), followed by selection in medium containing 2 μ g/ml blasticidin for NCI-H1703 cells and 0.5 μ g/ml blasticidin for NCI-H446 cells. TetR cells were subsequently infected with pLenti CMV/TO Hygro DEST virus (670-1) (Addgene plasmid no. 17293) containing the FSP1-GFP construct and were selected in medium containing 250 μ g/ml hygromycin. FSP1-GFP-expressing cells were enriched by fluorescence-activated cell sorting of the GFP-positive populations.

Generation of CRISPR-Cas9 genome-edited cell lines

For the CRISPR-Cas9 synthetic lethal screen, U-2 OS Tet-On lines stably expressing Cas9 were generated by infection with lentiCas9-Blast, a gift from F. Zhang (Addgene plasmid no. 52962) and cells were selected in medium containing 1 μ g/ml blasticidin. Active Cas9 expression was validated by flow cytometry analysis following infection with a self-cutting mCherry plasmid, which expresses mCherry and an sgRNA targeting the mCherry gene.

U-2 OS FSP1^{ko} lines were generated using CRISPR-Cas9 technology by transfection with pSpCas9(BB)-2A-Puro (PX459), a gift from F. Zhang (Addgene plasmid no. 48139), followed by selection in medium containing 1 μ g/ml puromycin and isolation of individual clones using cloning rings. U-2 OS COQ2^{ko} and FSP1^{ko} COQ2^{ko}, FSP1^{ko} ACSL4^{ko} and FSP1^{ko} NQO1^{ko} lines were generated by cotransfecting an FSP1^{ko} clonal line (FSP1 sgRNA guide 1, described in 'Plasmids') with PX459 plasmids encoding the appropriate guides, together with pcDNA3.1/Hygro(-) (Thermo Fisher Scientific) at a 20:1 w/w ratio, selection in medium containing 500 μ g/ml hygromycin, and isolation of individual clones using cloning rings. U-2 OS FSP1-HaloTag knock-in lines were generated by cotransfection of U-2 OS T-Rex Flp-In cells with the donor plasmid pUC57 (described in 'Plasmids') and PX459 encoding FSP1 sgRNA guide 3 at a 2:1 w/w ratio in medium containing 1 μ M SCR7 non-homologous end joining inhibitor (Xcess Biosciences) for 48 h, followed by selection in medium containing 1 μ g/ml puromycin.

NCI-H460 FSP1^{ko} lines were generated by infection with lentiCRISPR v2-Blast (Addgene plasmid no. 83489) virus, selection in medium containing 2 μ g/ml blasticidin and isolation of single clones using cloning rings. NCI-H460 GPX4^{ko} lines and FSP1^{ko} GPX4^{ko} lines were generated by infection with lentiCRISPR v2-Hygro (Addgene plasmid no. 98291)

virus, selection in medium containing 250 μ g/ml hygromycin, and isolation of single clones using cloning rings.

Plasmids

Cloning of all expression plasmids and the HaloTag donor plasmid was performed using restriction enzyme-independent fragment insertion by polymerase incomplete primer extension. To generate the FSP1-HaloTag knock-in donor plasmid, 800-base-pair homology arms flanking the FSP1 stop codon were amplified from U-2 OS genomic DNA and inserted in frame 5' and 3' to the linker-TEV-HaloTag sequence in pUC57 (a gift from R. Tjian). The protospacer adjacent motif site that corresponds to FSP1 sgRNA guide 3 was subsequently mutated in the donor sequence using mutagenesis primers to prevent cutting of the integrated donor sequence by Cas9. FSP1(WT)-GFP was generated by insertion of FSP1-GFP in pDEST47¹⁰ into pcDNA5/FRT/TO, and FSP1(G2A)-GFP and FSP1(E156A)-GFP were subsequently generated using site-directed mutagenesis. TOM20(SS)-FSP1(G2A)-GFP and LYN11-FSP1(G2A)-GFP were generated by insertion of the signal sequence of TOM20 and the first 11 amino acids of LYN kinase, respectively, at the N terminus of FSP1(G2A)-GFP. FSP1(G2A)-GFP-PLIN2 and FSP1(G2A)-GFP-Cb5 were generated by insertion of the full-length sequence for PLIN2 and amino acids 100-134 of cytochrome b5, respectively, at the C terminus of FSP1(G2A)-GFP. LYN11-mCherry-FRB was generated by replacement of CFP in LYN11-CFP-FRB²⁷ with the sequence for mCherry. BFP-Sec61 was a kind gift from G. Voeltz. FSP1-GFP in pLenti CMV/TO Hygro DEST (Addgene no. 17291) was generated by insertion of FSP1-GFP into pENTR1A, followed by Gateway recombination cloning (Thermo Fisher Scientific). NQO1-GFP was generated by PCR amplification of NQO1 from U-2 OS cDNA and insertion into pcDNA5/FRT/TO encoding GFP. LYN11-NQO1-GFP was generated by insertion of amino acids 1-11 of LYN at the N terminus of NQO1-GFP. For protein expression, FSP1(WT) and FSP1(E156A) lacking the ATG start codon were inserted into the pET-His6-TEV vector (Addgene plasmid no. 29653), C-terminal to the His6-TEV tag. LentiCas9-Blast was developed by the Zhang laboratory.

Plasmid transfections were performed in U-2 OS cells with Fugene6 (Promega) transfection reagent. Virus was produced by cotransfection of HEK293T cells with GAG, POL and pLenti expression plasmids at a 1:1:1 w/w/w ratio, using the X-tremeGENE HP (Roche) transfection reagent. Medium containing secreted virus was collected after 48 h and sterile-filtered.

CRISPR guide RNA (sgRNA) sequences targeting FSP1, ACSL4, NQO1, GPX4 and COQ2 were designed using the CRISPR design tool developed by the Zhang laboratory, available online (<http://crispr.mit.edu/>). The oligonucleotide sequences preceding the protospacer motif were: FSP1 guide 1, 5' caccgGAATCGGGAGCTCTGCACG 3'; FSP1 guide 2, 5' caccgTCCGATTCCACCGAGACCT 3'; FSP1 guide 3, 5' caccgTGAGGCAGTCTCACCTTGA 3'; ACSL4 guide 1, 5' caccgTGCAATCATCCATTCGGCCC 3'; ACSL4 guide 2, 5' caccgTGGTAGTGGACTCACTGCAC 3'; NQO1 guide 1, 5' caccgTTTGACGACTACCGACCA 3'; NQO1 guide 2, 5' caccgCAAGAGCACTGATCGTAC 3'; COQ2 guide, 5' caccgATGCTGGGCTCGCGAGCCGC 3'; and GPX4 guide, 5' caccgAGCCCCGCCGCGATGAGCCT 3'.

Nucleotides in lowercase show the overhangs introduced into oligonucleotides that are necessary for cloning into the BbsI restriction site of vector PX459 or BsmBI site of lentiCRISPR v2.

Chemicals and reagents

Reagents used in this study include: RSL3 (Cayman Chemical), Fer1 (Cayman Chemical), idebenone (Cayman Chemical), DFO (Cayman Chemical), doxycycline (Sigma), erastin2 (also known as compound 35MEW28) (synthesized by Acme), ML162 (Cayman Chemical), ZVAD(OMe)-FMK (Cayman Chemical), necrostatin-1 (Cayman Chemical), puromycin (Thermo Fisher Scientific), nutlin-3 (Cayman Chemical), CellEvent caspase-3/7 Green Detection Reagent (Thermo Fisher Scientific), etoposide (Sigma-Aldrich), rotenone (Sigma-Aldrich), blasticidin (Thermo

Article

Fisher Scientific), BODIPY 558/568 C12 (Thermo Fisher Scientific), BODIPY 493/503 (Thermo Fisher Scientific), BODIPY 581/591 C11 (Thermo Fisher Scientific), NMT inhibitor DDD85646 (Aobious), 4-CBA (Sigma-Aldrich), OptiPrep (Sigma-Aldrich), SYTOX Green Dead Cell Stain (Thermo Fisher Scientific), MitoTracker Green FM (Thermo Fisher Scientific), MitoTracker Orange CMTMRos (Thermo Fisher Scientific), CellMask Deep Red (Thermo Fisher Scientific), JF549 (kind gift from L. Lavis), oleate (Sigma-Aldrich), polybrene (Sigma-Aldrich), myristate (Sigma-Aldrich), YnMyr (Iris Biotech), AutoDOT (Abgent), DGAT1 inhibitor T863 (Sigma-Aldrich), DGAT2 inhibitor PF-06424439 (Sigma-Aldrich), SCR7 non-homologous end joining inhibitor (Xcess Biosciences), TAMRA-azide-PEG-biotin (BroadPharm), coenzyme Q₁ (Sigma-Aldrich), resazurin (Thermo Fisher Scientific) and NADH (Sigma-Aldrich). IKE was synthesized as previously described²¹.

Cell death analysis

Cells were plated in triplicate at a density of 2,000–3,000 cells per well in black 96-well plates (Corning) 48 h before start of imaging. To induce expression of FSP1, cells were treated with 10 ng/ml doxycycline at the time of plating. After 48 h, the medium was replaced with fresh medium containing 30 nM SYTOX Green Dead Cell Stain, doxycycline (if needed) and the indicated drugs. The plates were immediately transferred to an IncuCyte Zoom imaging system (Essen Bioscience) enclosed in an incubator set to 37 °C and 5% CO₂. Three images per well were captured in the green and phase channels every 1 or 2 h over a 48 h period, and the ratio of SYTOX Green-positive objects (dead cells) to phase objects (total cells) was quantified using Zoom image analysis software (Essen Bioscience). For each treatment condition, the SYTOX-to-phase-object ratio was plotted against the 48 h imaging interval, the AUC was calculated, and the average AUC was plotted as a function of drug concentration (for example, RSL3) using Prism (GraphPad). To calculate the half-maximal effective concentration (EC₅₀) values, the AUC curve was fit to a variable slope function comparing response to drug concentration. To quantify death in NQO1^{KO} cells, NQO1^{KO} FSP1^{KO} cells and lung cancer lines, SYTOX counts were used to calculate the AUC. To compare cell death between parental lung cell lines, the AUC values were normalized by the maximum value for each cell line.

For the 4-CBA treatment experiments, cells were treated with vehicle (1% v/v ethanol) or 3 mM 4-CBA 24 h after plating. Forty-eight hours after plating, the medium was replaced with fresh medium containing 4-CBA and the indicated drugs. For experiments comparing control to COQ2^{KO} cells, all cells were cultured in 200 μM uridine during imaging.

Western blotting

Cells were washed twice with PBS, lysed in 1% SDS, sonicated for 10 s and incubated for 5 min at 100 °C. Protein concentrations were determined using the bicinchoninic acid (BCA) protein assay (Thermo Fisher Scientific), and equal amounts of protein by weight were combined with 1× Laemmli buffer, separated on 4–20% polyacrylamide gradient gels (Bio-Rad Laboratories) and transferred onto nitrocellulose membranes (Bio-Rad Laboratories). Membranes were washed in PBS with 0.1% Tween-20 (PBST) and blocked in PBST containing 5% (w/v) dried milk for 30 min. Membranes were incubated for 24 h in PBST containing 5% bovine serum albumin (BSA) (Sigma Aldrich) and primary antibodies. After washing with PBST, membranes were incubated at room temperature for 30 min in 5% BSA and PBST containing fluorescent secondary antibodies. Immunoblots were imaged on a LI-COR imager (LI-COR Biosciences).

The following blotting reagents and antibodies were used: anti-PLIN2 (Abgent), anti-AIFM2 (Proteintech Group and Santa Cruz Biotechnology), anti-α-tubulin (Cell Signaling Technology and Santa Cruz Biotechnology), anti-GPX4 (Abcam), anti-ACSL4 (Sigma-Aldrich), anti-GFP (Proteintech Group), anti-NQO1 (Proteintech Group), anti-GAPDH (EMD Millipore), anti-RAS (Cell Biolabs), anti-MDR1 (Cell Signaling Technology), anti-p21 (Cell Signaling Technology), anti-rabbit IRDye800

conjugated secondary (LI-COR Biosciences) and anti-mouse Alexa Fluor 680 conjugated secondary (Invitrogen).

Fluorescence microscopy

For fluorescence microscopy of PLIN2 and FSP1-GFP in fixed cells, cells grown on coverslips were treated with 200 μM oleate-BSA complex for 24 h, washed 3× with PBS, fixed for 15 min in PBS containing 4% (w/v) paraformaldehyde and washed 3× again with PBS. Cells were permeabilized for 15 min with blocking solution (1% BSA and PBS) containing 0.01% digitonin, washed 3× and incubated in blocking solution for an additional 15 min. Cells were incubated with anti-PLIN2 antibody in blocking solution (1:500 dilution) for 2 h at room temperature, washed 3× and incubated for 1 h in blocking solution containing anti-rabbit secondary antibody conjugated to Alexa Fluor 594 (1:500 dilution) (Thermo Fisher Scientific). After additional 3× washes, coverslips were mounted on glass slides using Fluoromount G (Southern Biotech). For fluorescence microscopy of FSP1-GFP and LYN-mCherry-FRB, cells were fixed in PBS containing 4% (w/v) paraformaldehyde and washed 3× with PBS before mounting.

For live-cell microscopy, cells were grown in 4-well or 8-well Laboratory-Tek II Chambered Coverglass (Thermo Fisher Scientific) imaging chambers. To image lipid droplets, cells were incubated for 24 h with 1 μM BODIPY 558/568 C12 (Thermo Fisher Scientific) or treated with 100 μM AutoDOT before imaging. To image the cell membrane, cells were incubated with 5 μg/ml CellMask Deep Red for 30 min, and the medium was replaced before imaging. To image mitochondria, cells were incubated with 100 nM MitoTracker Orange CMTMRos or MitoTracker Green FM for 15 min. For imaging that required prior transfection, cells were transiently transfected with the indicated plasmids in 6-well plates using Fugene6, incubated for 48 h and seeded in Laboratory-Tek II chambers before imaging. To image FSP1-HaloTag, cells were incubated with 100 nM JF549 dye for 30 min, washed 3× with PBS and imaged in fresh medium.

Cells were imaged using a Deltavision Elite widefield epifluorescence deconvolution microscope (GE Healthcare) equipped with a 60× oil immersion objective (Olympus), using DAPI, FITC, Tx-Red and Cy5 filters. For live-cell microscopy, cells were imaged in an enclosure heated to 37 °C and exposed to a continuous perfusion of a gas mixture containing 5% CO₂, 21% O₂ and 74% N₂ (BioBlend, Praxair). Z-stacks of 0.2-μm slices totalling 4–6 μm in thickness were acquired for deconvolution using SoftWoRx software (GE Life Sciences). Single deconvolved slices for each channel were analysed and merged using ImageJ (<http://imagej.nih.gov/ij/>).

Lipid droplet fractionation

Ten 15-cm plates of U-2 OS cells expressing inducible FSP1-GFP were induced with 10 ng/ml doxycycline for 48 h. Cells were collected by scraping into PBS and centrifuged for 10 min at 500g. Cell pellets were resuspended in cold hypotonic lysis medium (HLM, 20 mM Tris-HCl pH 7.4 and 1 mM EDTA) supplemented with 1× cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail (Sigma-Aldrich), incubated on ice for 10 min, dounced using 80× strokes and centrifuged at 1,000g for 10 min. The supernatant was subsequently transferred to Ultra-Clear ultracentrifuge tubes (Beckman-Coulter), diluted with 60% sucrose and HLM to a final concentration of 20% sucrose and HLM, and overlaid by 4 ml of 5% sucrose and HLM followed by 4 ml of HLM. Overlaid samples were centrifuged for 30 min at 15,000g in an ultracentrifuge using a SW41 swinging bucket rotor (Beckman-Coulter). Buoyant fractions were collected using a tube slicer (Beckman-Coulter), additional fractions were pipetted from the top of the sucrose gradient in 1-ml increments, and pellets were resuspended in 1 ml HLM. One hundred microlitres of 10% SDS was added to each fraction, yielding a final concentration of 1% SDS. Samples were then sonicated for 15 s and incubated for 10 min at 65 °C. Buoyant fractions were incubated at 37 °C for 1 h and sonicated every 20 min, followed by a final incubation at 65 °C for 10 min.

Plasma-membrane fractionation

Plasma-membrane subdomains were separated using a continuous OptiPrep gradient as previously described²⁸. Six 15-cm plates of cells expressing inducible FSP1–GFP were incubated with 10 ng/ml doxycycline for 48 h and collected by scraping into PBS, centrifuged for 10 min at 500g and resuspended in 1 ml of base buffer (20 mM Tris-HCl pH 7.8 and 250 mM sucrose) supplemented with 1 mM MgCl₂, 1 mM CaCl₂, and 1× cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail. Cells were passed 40× through a 1.5'' 22-gauge needle and centrifuged at 1,000g for 10 min. The supernatant was retained, and the pellet was resuspended in an additional 1 ml base buffer containing 1 mM MgCl₂ and 1 mM CaCl₂. The resuspended pellet was passed 40× through a 22-gauge needle, centrifuged at 1,000g for 10 min and the supernatant was combined with 1 ml of supernatant from the previous step to make 2 ml in total. OptiPrep mixing solution was prepared by combining 60% OptiPrep stock solution with buffer containing 120 mM Tris-HCl pH 7.8 and 250 mM sucrose in a 5:1 v/v ratio. Two millilitres OptiPrep mixing solution was combined with 2 ml of sample supernatant from the previous centrifugation steps to yield 4 ml of a sample containing 25% OptiPrep. This OptiPrep-mixed sample was gently pipetted under 8 ml of a continuous 5–20% OptiPrep gradient prepared in base buffer in an UltraClear tube. The loaded sample was subsequently centrifuged for 90 min at 52,000g at 4 °C using a SW41 swinging bucket rotor. After centrifugation, individual 0.67-ml fractions were collected by pipetting from the top of the gradient and analysed by western blot. The plasma-membrane-localized proteins RAS and MDRI were used as markers of plasma-membrane fractions.

CRISPR–Cas9 synthetic lethal screen

The CRISPR–Cas9 screen was performed as previously described²⁹. The ‘Apoptosis and Cancer’ sublibrary of sgRNAs²⁹ comprising 31,324 elements—including 29,824 sgRNAs targeting 3,015 genes (about 10 sgRNAs per gene) and 1,500 negative-control sgRNAs—was used. To generate lentiviral particles, the sublibrary was co-transfected with third-generation lentiviral packaging plasmids (pVSVG, pRSV and pMDL) into HEK293T cells. Medium containing lentivirus was collected 48 and 72 h after transfection, combined, filtered and then used to infect about 2.1×10^7 U-2 OS Tet-On cells stably expressing Cas9. After 72 h of growth, infected cells were selected in medium containing 1 µg/ml puromycin until over 90% cells were mCherry-positive. Cells were then re-seeded in 500-cm² plates (about 8×10^6 cells per plate) and recovered in medium lacking puromycin for 24 h. For the screen, a total of about 3.2×10^7 cells (that is, about 1,000-fold library coverage) were treated with either DMSO or 0.5 µM RSL3 for 5 days. Cells were then trypsinized, collected by centrifugation at 1,000g, washed twice with PBS and pellets were frozen at -80 °C. Genomic DNA was extracted using the QIAamp DNA Blood Maxi Kit (QIAGEN) according to the manufacturer's instructions. sgRNA sequence libraries were prepared from genomic DNA by two rounds of PCR using the Herculase II Fusion DNA Polymerase (Agilent). sgRNA sequences were amplified by the primers oMCB1562 and oMCB1563 and then indexed using the Illumina TruSeq LT adaptor sequences AD006 (GCCAAT; DMSO) or AD012 (CTTGTA; RSL-3) for downstream deep sequencing analysis. PCR products were separated on a 2% tris-borate-EDTA (TBE)-agarose gel, purified using the QIAquick Gel Extraction Kit (Qiagen) and assessed for quality using a Fragment Analyzer (Agilent). PCR amplicons from each sample were pooled in a 1:1 ratio based on their concentrations as determined by Qubit Fluorometric Quantification. sgRNA sequences were analysed by deep sequencing using the primer oMCB1672 on an Illumina MiSeq instrument at the Oklahoma Medical Research Foundation. Sequence reads were aligned to the sgRNA reference library using Bowtie software. For each gene, a gene effect and score (likely maximum effect size and score), and *P* value

were calculated using the castLE statistical framework as previously described²⁹.

Click chemistry and in-gel fluorescence

To analyse myristoylated proteins in buoyant fractions enriched in lipid droplets, 20 15-cm plates of U-2 OS cells were incubated with 100 µM myristic acid or 100 µM YnMyr for 48 h. Buoyant fractions were isolated by sucrose gradient fractionation as described in ‘Lipid droplet fractionation’, combined with SDS (1% final concentration) and dialysed into a 0.1% SDS and PBS solution. A click mixture was prepared by adding reagents in the following order, and by vortexing after the addition of each reagent: 10 µl of 10 mM TAMRA–azide–PEG–biotin (BroadPharm), 20 µl of 50 mM copper (II) sulfate, 20 µl of 50 mM tris(2-carboxyethyl)phosphine, 10 µl of 10 mM tris(benzyltriazolylmethyl) amine. Sixty microlitres of click mixture was then added to 1 ml of the dialysed samples. The samples were then vortexed and incubated for 1 h at room temperature. One millilitre of cold methanol containing 10 mM EDTA was added to each sample, and the samples were briefly vortexed and stored at -80 °C overnight. The following day, the samples were centrifuged at 17,000g at 4 °C for 30 min to pellet precipitated proteins. Pellets were washed twice with 1 ml cold methanol, dried in a speed-vacuum centrifuge under medium heat, and resuspended in 80 µl PBS containing 1% SDS. Once dissolved, proteins were resuspended in 1× Laemmli loading buffer and analysed by SDS–PAGE. To visualize proteins using fluorescence, the gel was washed 3× with milliQ water and imaged using a ChemiDoc XRS+ Imaging System (Bio-Rad Laboratories).

To analyse myristoylated proteins in whole-cell lysates, U-2 OS cells were incubated with 100 µM myristic acid or 100 µM YnMyr for 48 h. Cells were washed twice with PBS and lysed in buffer containing 1% SDS and PBS and 1× EDTA-free complete protease inhibitor. Equal amounts of protein by weight were diluted to 0.1% SDS and PBS and subjected to click chemistry with TAMRA–azide–PEG–biotin.

Enrichment of *N*-myristoylated proteins

YnMr-labelled proteins in cell lysates were conjugated to TAMRA–azide–PEG–biotin using click chemistry as described in ‘Click chemistry and in-gel fluorescence’. After protein precipitation in cold methanol, the pellet was resuspended in 80 µl of 1% SDS and PBS. Once the pellet was completely dissolved, 65 µl was diluted 10-fold with PBST. Fifteen microlitres of Streptavidin Agarose Resin (Thermo Fisher Scientific) was washed 3× with PBST. The diluted sample was added to the bead resin and rotated for 3 h at room temperature. Beads were washed 5× with PBST and bound proteins were eluted by boiling the beads for 5 min in 2× Laemmli buffer containing 2 mM biotin.

Lipidomic profiling using liquid chromatography–tandem mass spectrometry

Cas9^{etl} and FSP1^{KO} U-2 OS cells grown in 10-cm plates were scraped into PBS, centrifuged at 500g for 5 min, and processed as previously described²⁴. After addition of internal standards (10 nmol of dodecyl-glycerol and 10 nmol of pentadecanoic acid), lipids were extracted in a 4 ml solution of 2:1:1 chloroform:methanol:PBS. The organic and aqueous layers were separated by centrifugation at 1,000g for 5 min. Following the collection of the organic layer, the remaining organic material in the aqueous layer was acidified by addition of 0.1% formic acid and re-extracted with 2 ml of chloroform. Extracts were combined, dried down under a stream of nitrogen and then resolubilized in 120 µl of chloroform. Ten microlitres of sample was analysed by single reaction monitoring-based liquid chromatography–mass spectrometry. Liquid chromatography separation was performed using a Luna reverse-phase C5 column, and mass spectrometry analysis was performed using an Agilent 6400 triple quadrupole (QQQ)–liquid chromatography–mass spectrometry instrument. Metabolites were quantified by integrating the AUC, and the values were normalized to the internal standards.

Glutathione measurements

The day before the experiment, 2×10^5 Cas9^{ctrl} and FSP1^{KO} U-2 OS cells per well were seeded into 6-well dishes. Cells were treated with DMSO (vehicle), erastin2 (1 μ M) for 6 h or RSL3 (250 nM) for 1 h. Cells were collected by scraping and prepared for measurement of glutathione (GSH + GSSG) using the Cayman Chemical Glutathione Assay Kit (Cayman Chemical) according to the manufacturer's protocol. The GSH and GSSG concentrations were calculated using a standard curve and normalized to the total protein level in each sample. Three independent biological replicates were performed for each condition.

BODIPY 581/591 C11 analysis

The day before the experiment, 2×10^5 U-2 OS cells per well were seeded into 6-well dishes containing a 22-mm² glass coverslip in each well. Cells were treated with DMSO (vehicle) or RSL3 (250 nM) for 75 min. At the end of the treatment, the treatment medium was removed and cells were washed once with HBSS. Cells were then labelled in 1 ml HBSS containing 5 μ M BODIPY 581/591 C11 and incubated at 37 °C for 10 min. The label mixture was removed and 1 ml of fresh HBSS was added to the cells. The cover slip was transferred to a glass microscope slide onto which 25 μ l of fresh HBSS had been applied. Confocal imaging and quantification of BODIPY 581/591 C11 were performed as previously described¹¹ on two independent biological replicates per treatment. Using ImageJ, each nucleus was attributed two regions of interest (ROI), one perinuclear and one plasma membrane-localized. Red and green fluorescence values were quantified for each ROI and corrected for background by subtracting the red or green fluorescence in cell-free areas. The BODIPY 581/591 C11 value was calculated as the ratio of the green fluorescence (which indicates oxidized probe) to total (green + red, which indicates total reduced plus oxidized probe) fluorescence.

Tumour xenograft growth studies

For Fer1 withdrawal experiments, tumour xenografts were established by injection of GPX4^{KO} and GPX4^{KO} FSP1^{KO} H460 cells into the flank of male C.B17 SCID mice, 6 weeks of age (Taconic Farms) ($n = 8$). In brief, cells were washed with PBS, trypsinized and collected in serum-containing medium. Collected cells were then washed with serum-free medium once and resuspended in serum-free medium at a concentration of 2×10^4 cells/ μ l. One hundred microlitres of cells (2×10^6 cells) were injected per mouse. Fer1 was prepared at a concentration of 0.2 mg/ml in 18:1:1 v/v PBS:ethanol:PEG40. Mice were injected intraperitoneally with Fer1 daily (2 mg kg⁻¹ body weight), and tumour size was measured using callipers. Fer1 injections were discontinued in 1 set of mice 5 days after cell injection, and tumour size was measured once every 2 days in each mouse for an additional 17 days. Mice not included in the analysis included mice that were killed early owing to sickness ($n = 1$ of GPX4^{KO} (+) Fer1) and mice with tumours that were determined to be outliers according to the statistical test described in 'Statistical analysis and reproducibility' ($n = 1$ of GPX4^{KO} (-) Fer1 and $n = 1$ of GPX4^{KO} FSP1^{KO} (-) Fer1). The number of mice represented in the final analysis was GPX4^{KO} (-) Fer1 ($n = 7$), GPX4^{KO} (+) Fer1 ($n = 7$), GPX4^{KO} FSP1^{KO} (-) Fer1 ($n = 7$), and GPX4^{KO} FSP1^{KO} (+) Fer1 ($n = 8$).

For IKE injection experiments, IKE was resuspended at 4 mg/ml in an HBSS pH 4 solution containing 4% DMSO, 2% ethanol and 4% PEG40. To prepare this solution, 24 mg of IKE was dissolved in 200 μ l DMSO, and 100 μ l ethanol and 250 μ l PEG40 were sequentially added. This mixture was added to 5.4 ml of HBSS pH 4 (Gibco) and sterile-filtered. Tumour xenografts were established by injection of Cas9^{ctrl} ($n = 8$) or FSP1^{KO} H460 ($n = 8$) cells. After 10 days, each group of mice ($n = 8$) was injected daily with vehicle or 40 mg kg⁻¹ IKE (250 μ l total volume), and the fold change in tumour size was measured over a 24-day period. Mice not included in the analysis included mice that were killed early owing to the development of exceedingly large tumours ($n = 1$ of Cas9^{ctrl} (+) IKE,

$n = 3$ of Cas9^{ctrl} (-) IKE, $n = 3$ of FSP1^{KO} (+) IKE and $n = 1$ of FSP1^{KO} (-) IKE), mice in which tumours did not initiate ($n = 2$ of Cas9^{KO} (+) IKE), and mice with tumours that were determined to be statistical outliers according to the test described in 'Statistical analysis and reproducibility' ($n = 1$ of Cas9^{ctrl} (+) IKE, $n = 1$ of Cas9^{ctrl} (-) IKE, $n = 1$ of FSP1^{KO} (+) IKE). The number of mice represented in the final analysis was Cas9^{ctrl} (-) IKE ($n = 4$), Cas9^{ctrl} (+) IKE ($n = 4$), FSP1^{KO} (-) IKE ($n = 7$) and FSP1^{KO} (+) IKE ($n = 4$).

No statistical tests were used to calculate sample size. Sample size was $n = 8$ for each treatment group to account for differences in tumour formation and growth, and to ensure recovery of a sufficient quantity of mice with tumours of approved size at each time point of the study. Following injection of H460 cells, the mice were randomly assigned into two treatment groups for the Fer1 withdrawal experiments and into two treatment groups for the IKE injection experiments. Fold change in tumour volume was statistically analysed using the unpaired, two-way *t*-test. Blinding was not possible because the experiments were performed by a single researcher. All mouse experiments were conducted in accordance with the guidelines of the Institutional Animal Care and Use Committees (IACUC) of the University of California, Berkeley. Animals were euthanized when the xenograft tumour size reached two centimetres in any two dimensions. No mouse exhibited severe loss of body weight (>15%) or evidence of infections or wounds.

TIRF microscopy

Cells were imaged at room temperature using a Nikon Ti-E inverted microscope (Nikon Instruments) outfitted with a TIRF 60 \times /1.49 NA oil objective, an Andor Laser Combiner and an electron-multiplying charge-coupled device camera (iXon ULTRA 897BV; Andor Technology). Samples were excited with a 488-nm laser line, and emission was collected through a single band-pass filter centred on 510 nm. All images were acquired using iQ3 acquisition software (Andor Technology). The depth of the evanescent field was approximately 150 nm.

CoQ measurements

CoQ measurements were performed as previously described³⁰. To simultaneously measure the reduced and oxidized form of CoQ, a cold butylhydroxytoluene (BHT) solution was added to prevent auto-oxidation at the beginning of sample extraction. One hundred microlitres of a cold BHT-in-propanol solution (5 mg/ml) and 600 μ l of cold 1-propanol were added to each tube containing cells in the frozen state. Immediately after this, the mixture was subjected to sonication for 2 min. Subsequently, 100 μ l of cold coenzyme Q₉ solution (2 μ g/ml), which was used as internal standard, was added, and the mixture was vortex-mixed for 1 min. It was then centrifuged for 10 min at 3,500 rpm and 1 °C, and the propanol organic supernatant layer was transferred to an autosampler vial. One-hundred-microlitre aliquots of the 1-propanol extract were immediately analysed, and the reduced and oxidized CoQ levels were determined using high-performance liquid chromatography (HPLC). HPLC analysis was performed using an automated Hitachi Chromaster system equipped with a Model 5110 quaternary pump, Model 5210 autosampler, Model 5310 column oven and ESA CouloChem III detector. The EZChrom Elite software (Agilent) was used for monitoring output signal and processing the results. The analytical column was a 150-mm \times 4.6-mm C18 column with 5- μ m spherical particles connected to a Security Guard equipped with a C18 cartridge (4-mm \times 3-mm).

Apoptosis activation assay

Cells grown in 6-cm plates were washed with PBS, trypsinized and centrifuged for 5 min at 500g. Cell pellets were resuspended in PBS containing 5% FBS and 5 μ M CellEvent caspase-3/7 Green Detection Reagent and were incubated for 30 min at 37 °C. Cells were analysed on a LSRFortessa (Becton Dickinson) flow cytometer, and the raw data were processed using the FlowJo software package (TreeStar). Apoptotic cells were gated using the same forward scatter threshold across all samples, and FITC fluorescence of the gated populations was determined.

Protein purification and activity assays

Expression vectors were transformed into Rosetta DE3 competent cells (EMD Millipore) and LB cultures were inoculated for overnight growth at 37 °C while shaking. The following day, the cultures were diluted 1:100 into 500 ml of LB and allowed to grow to an optical density at 600 nm (OD_{600}) of 0.5, at which point the incubator was set to start cooling to 20 °C. The cultures were grown further to an OD_{600} of 0.7 and induced with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) overnight. Bacterial pellets were resuspended in 2 ml of cold lysis buffer containing 50 mM potassium phosphate pH 8.0, 300 mM potassium chloride, and 30 mM imidazole, supplemented with 1× cComplete, Mini, EDTA-free Protease Inhibitor Cocktail. The resuspended cells were sonicated 5× on ice at 50% power for 10 s, with 2-min incubations on ice in between sonications, and were centrifuged at 20,000g for 15 min at 4 °C. The supernatant was combined with 200 μ l of Ni-NTA agarose beads (Thermo Fisher Scientific) washed 3× with lysis buffer, and the supernatant–bead mixture was rotated for 1 h at 4 °C. The beads were subsequently washed 5× with cold lysis buffer, and bound proteins were eluted by incubating beads for 15 min in 500 μ l of cold lysis buffer containing 250 mM imidazole while rotating. The eluted proteins were dialysed into PBS containing 10% glycerol and snap-frozen in liquid N_2 . Protein concentration was determined by measuring the absorbance at 280 nm.

To measure NADH oxidation kinetics, recombinant FSP1 was combined with 500 μ M NADH and 200 μ M coenzyme Q_1 in a total volume of 100 μ l PBS. A reduction in absorbance at 340 nm, corresponding to NADH oxidation, was determined over the course of 1 h. To measure resazurin reduction kinetics, recombinant FSP1 was combined with 500 μ M NADH and 500 μ M resazurin in a total volume of 100 μ l PBS. Fluorescence (emission at 590 nm) corresponding to reduced resazurin was determined over the course of 1 h. All measurements were taken using a SpectraMax i3 Multi-Mode Platform plate reader (Molecular Devices).

Analysis of the CTRP dataset

Data for significant correlations between *FSP1* gene expression and resistance to RSL3, ML162 and ML210 were downloaded from the CTRP v2 website¹⁸. Data for non-haematopoietic cancer cells was extracted from the v21.data.gex_global_analysis.txt table and plotted using Prism.

Statistical analysis and reproducibility

All figures, including western blots, dose–response curves and enzymatic activity assay panels are representative of two biological replicates unless stated otherwise. Images are representative of at least $n = 10$ imaged cells. P values for pairwise comparisons were calculated using the two tailed t -test. For comparison across multiple experimental groups, P values were calculated using one-way ANOVA, and adjusted using Bonferroni correction for multiple comparisons. For Fig. 4a, b and Extended Data Fig. 10a, b, the normalized z -scored Pearson correlation coefficients were obtained from CTRP v2 (<https://portals.broadinstitute.org/ctrp/>). For xenograft experiments, all mice were randomized following tumour-cell injection into treatment groups. Outliers were identified using the Grubbs method, and were removed from analyses. To compare between groups of mice in each time point, P values were calculated using the unpaired, two way t -test.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data that support the conclusions in this manuscript are available from the corresponding author upon reasonable request. Raw data for Fig. 1 are provided in Supplementary Table 1. Raw data for Fig. 3 are provided in Supplementary Table 3. Raw data for Fig. 4 are provided in Supplementary Table 4, and are publicly available from the Cancer Cell Line Encyclopedia (<https://portals.broadinstitute.org/ccle>) and CTRP databases.

Code availability

The castLE statistical framework software for analysis of data from the CRISPR screen can be accessed at www.bitbucket.org/dmorgens/castle/. Bowtie software can be accessed at www.bowtie-bio.sourceforge.net/bowtie2/index.shtml. MATLAB image analysis software to analyse lipid droplet distributions can be obtained at www.droplet-proteome.org.

27. Inoue, T., Heo, W. D., Grimley, J. S., Wandless, T. J. & Meyer, T. An inducible translocation strategy to rapidly activate and inhibit small GTPase signaling pathways. *Nat. Methods* **2**, 415–418 (2005).
28. Macdonald, J. L. & Pike, L. J. A simplified method for the preparation of detergent-free lipid rafts. *J. Lipid Res.* **46**, 1061–1067 (2005).
29. Morgens, D. W. et al. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017).
30. Tang, P. H., Miles, M. V., DeGrauw, A., Hershey, A. & Pesce, A. HPLC analysis of reduced and oxidized coenzyme Q_{10} in human plasma. *Clin. Chem.* **47**, 256–265 (2001).

Acknowledgements This research was supported by grants from the National Institutes of Health (R01GM112948 to J.A.O., 1R01GM122923 to S.J.D., P42 ES004705 to D.K.N. and 1DP2CA195761-01 to R.Z.). J.A.O. is a Chan Zuckerberg Biohub investigator. D.K.N. was supported by a Cancer Research ASPIRE award from the Mark Foundation. P.H.T. was supported by the Internal Research Fund of the Division of Pathology and Laboratory Medicine, Cincinnati Children's Hospital Medical Center. We thank P.-J. Ko (Stanford) for assistance with confocal imaging, and D. Leto (Stanford) and R. Kopito (Stanford) for helpful discussions.

Author contributions K.B. and J.A.O. conceived the project and designed the experiments. J.A.O. and K.B. wrote the manuscript. All authors read and edited the manuscript. K.B. performed the majority of the experiments. Z.L. and M.A.R. performed and analysed the CRISPR screen with guidance from M.C.B. K.B. prepared samples and R.Z. performed the TIRF microscopy, B.F. performed the lipidomics, and P.H.T. measured CoQ levels and redox state. J.H. performed the click chemistry myristoylation experiments. S.J.D. and L.M. performed the glutathione measurements and C11 experiments. J.M.H. generated the overexpression and knockout lung cancer lines and analysed ferroptosis in these lines. D.K.N., J.M.H., B.F., and M.A.R. performed the xenograft experiments. T.J.M. and B.T. synthesized IKE.

Competing interests S.J.D. is a member of the scientific advisory board for Ferro Therapeutics.

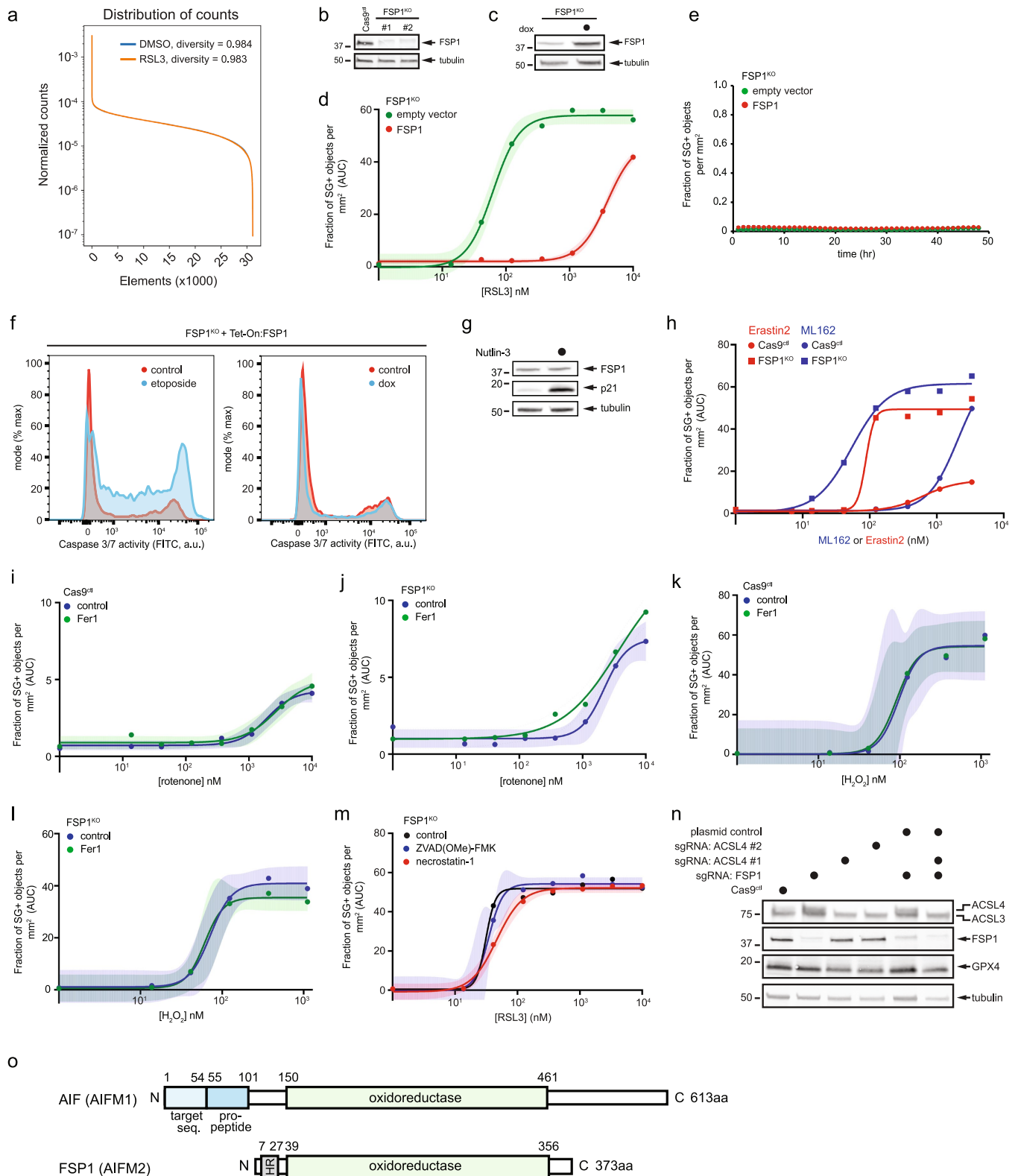
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1705-2>.

Correspondence and requests for materials should be addressed to J.A.O.

Peer review information Nature thanks Kivanc Birsoy, Navdeep S. Chandel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

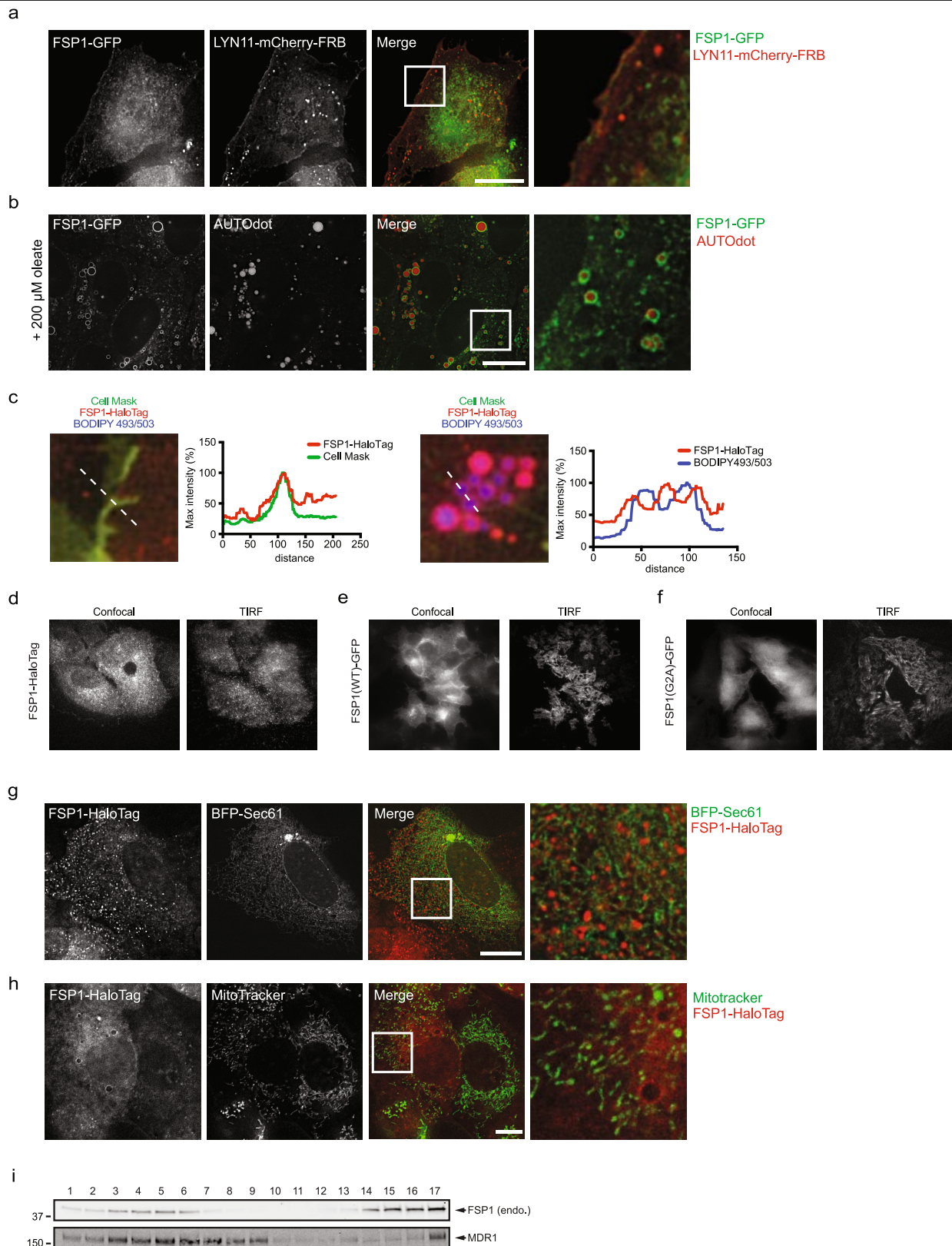
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Synthetic lethal screen coverage and validation.

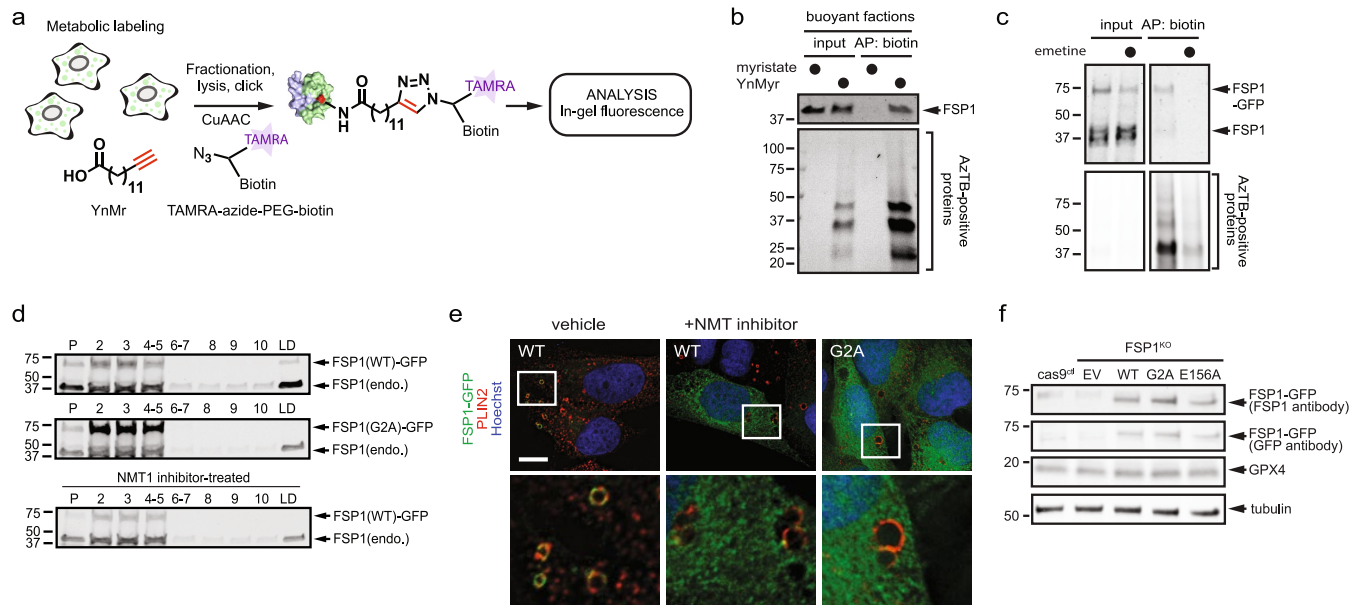
a, Distribution of counts across all sgRNA elements from the CRISPR-Cas9 screen. **b**, Western blot of control and FSP1^{KO} cells. **c**, Western blot analysis (c) and dose response of RSL3-induced death (d) of FSP1^{KO} cells that express doxycycline-inducible, untagged FSP1. **e**, Time-lapse analysis of cell death of FSP1^{KO} cells that express inducible, untagged FSP1. **f**, Flow cytometric analysis of caspase 3/7 activity in FSP1^{KO} cells that express inducible, untagged FSP1, treated with doxycycline for 48 h. As a positive control, non-induced cells were treated with 50 μ M etoposide for 24 h before analysis. **g**, Western blot analysis of lysates from control cells treated with 10 μ M nutlin-3 for 48 h. **h**, Dose

response of ML162 and erastin2-induced cell death. **i**, **j**, Dose response of rotenone-induced death of control (i) and FSP1^{KO} (j) cells. **k**, **l**, Dose response of hydrogen-peroxide-induced death of control (k) and FSP1^{KO} (l) cells. **m**, Dose response of RSL3-induced cell death in the presence of inhibitors of apoptosis (ZVAD(OMe)-FMK, 10 μ M) and necroptosis (necrostatin-1, 1 μ M). **n**, Western blot analysis of lysates from ACSL4^{KO} and FSP1^{KO} ACSL4^{KO} cells. **o**, Schematic of domains present in AIF and FSP1. In **d**, **i**–**m**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. Panels are representative of two biological replicates, except panels **c**–**e** and **k**, **l**, which show single experiments.



Extended Data Fig. 2 | Subcellular distribution of FSP1. **a**, Inducible FSP1-GFP cells were transiently transfected with LYN11-mCherry-FRB for 24 h, induced with doxycycline for 48 h and fixed before imaging. **b**, FSP1-GFP cells were treated with 200 μ M oleate for 24 h to induce lipid droplets and treated with 100 μ M AutoDOT to label lipid droplets before imaging. **c**, Line intensity plots showing colocalization between FSP1-HaloTag and organelle markers. **d-f**, Confocal and TIRF microscopy of FSP1-HaloTag (**d**), and inducible

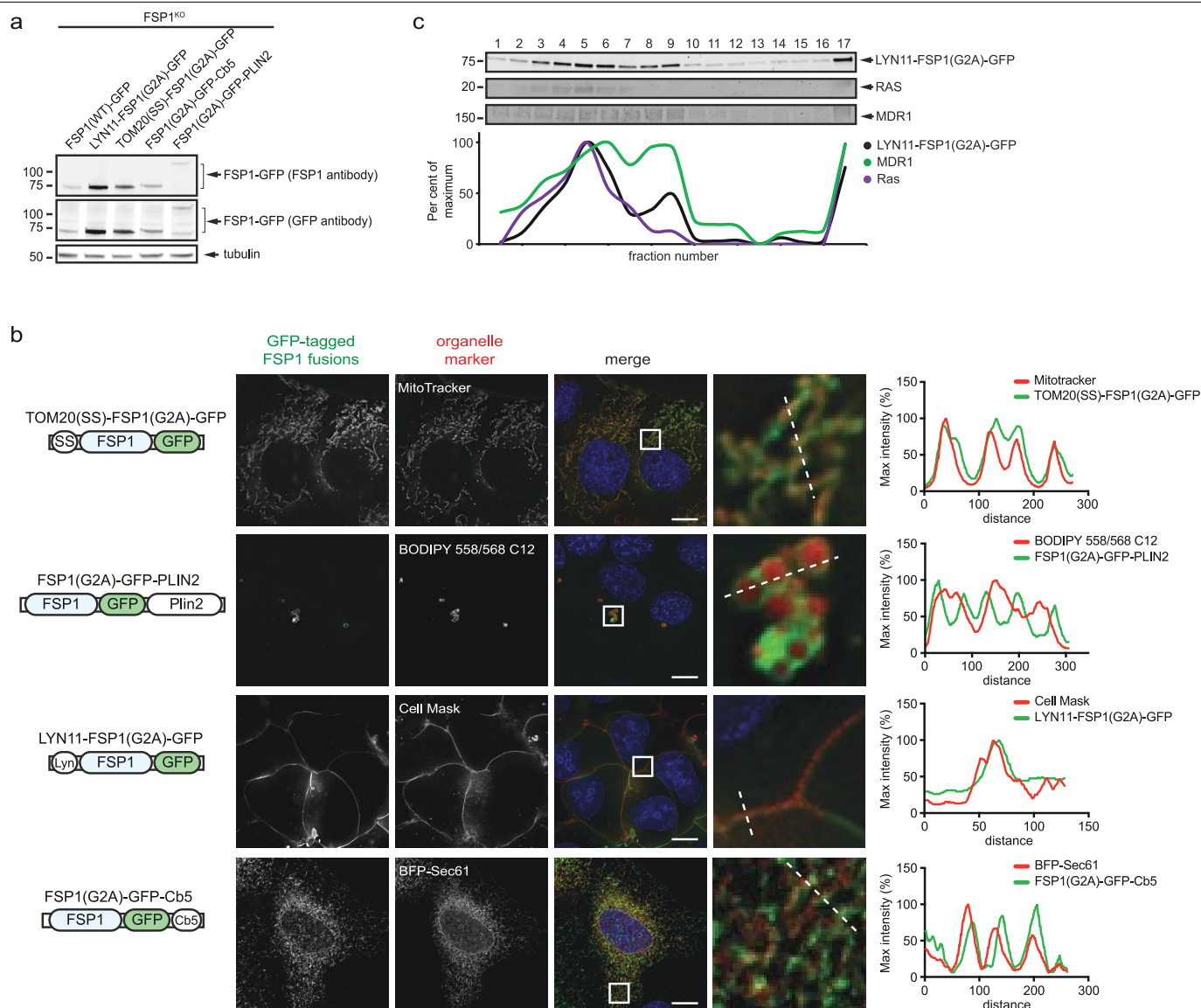
FSP1(WT)-GFP (**e**) and FSP1(G2A)-GFP (**f**) cells. **g**, FSP1-HaloTag cells were transiently transfected with BFP-Sec61 for 48 h before imaging to label the endoplasmic reticulum. **h**, FSP1-HaloTag cells were incubated with 100 nM MitoTracker Green FM to label mitochondria. **i**, Plasma-membrane subdomains from control cells were enriched by OptiPrep gradient centrifugation. Endo., endogenous FSP1. Western blot is representative of two biological replicates. Images are representative of at least $n = 10$ imaged cells. Scale bars, 10 μ m.



Extended Data Fig. 3 | Myristoylation and lipid droplet localization of FSP1.

a, Schematic showing the procedure for metabolic labelling of cells with the myristate-alkyne YnMyr and conjugation of YnMyr-labelled proteins with TAMRA-azide-PEG-biotin using click chemistry. **b**, Analysis of FSP1 myristoylation in buoyant fractions enriched in lipid droplets, by streptavidin enrichment of YnMyr-labelled proteins, click chemistry and SDS-PAGE. Cells were treated with 200 μ M oleate to induce lipid droplets and with 100 μ M YnMyr or 100 μ M myristate for 24 h. **c**, FSP1-GFP was induced with doxycycline for 24 h and cells were incubated with 100 μ M YnMyr for an additional 24 h to label proteins in the presence or absence of 75 μ M emetine. YnMyr-labelled

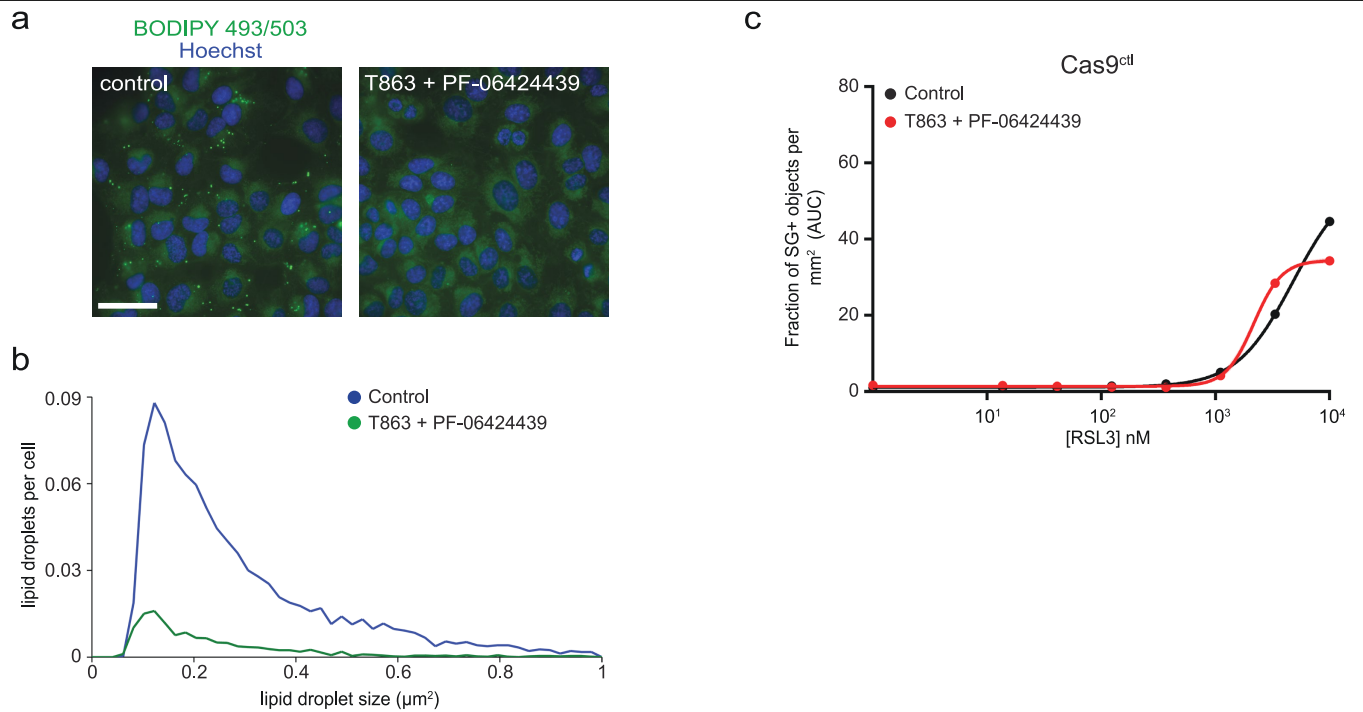
proteins were affinity-purified and analysed by click chemistry and SDS-PAGE. **d**, Buoyant fractions enriched in lipid droplets, from cells expressing inducible FSP1-GFP, were isolated by sucrose gradient fractionation and analysed by western blot. Endo., endogenous FSP1. **e**, Inducible FSP1-GFP cells were treated with 200 μ M oleate in the presence or absence of 10 μ M NMT inhibitor, fixed and stained with anti-PLIN2 antibody before imaging. Images are representative of at least $n = 10$ imaged cells. Scale bar, 10 μ m. **f**, Western blot analysis of FSP1^{KO} cells induced for 48 h with doxycycline to express the indicated proteins. All panels are representative of two biological replicates.



Extended Data Fig. 4 | Targeting of FSP1 to subcellular compartments.

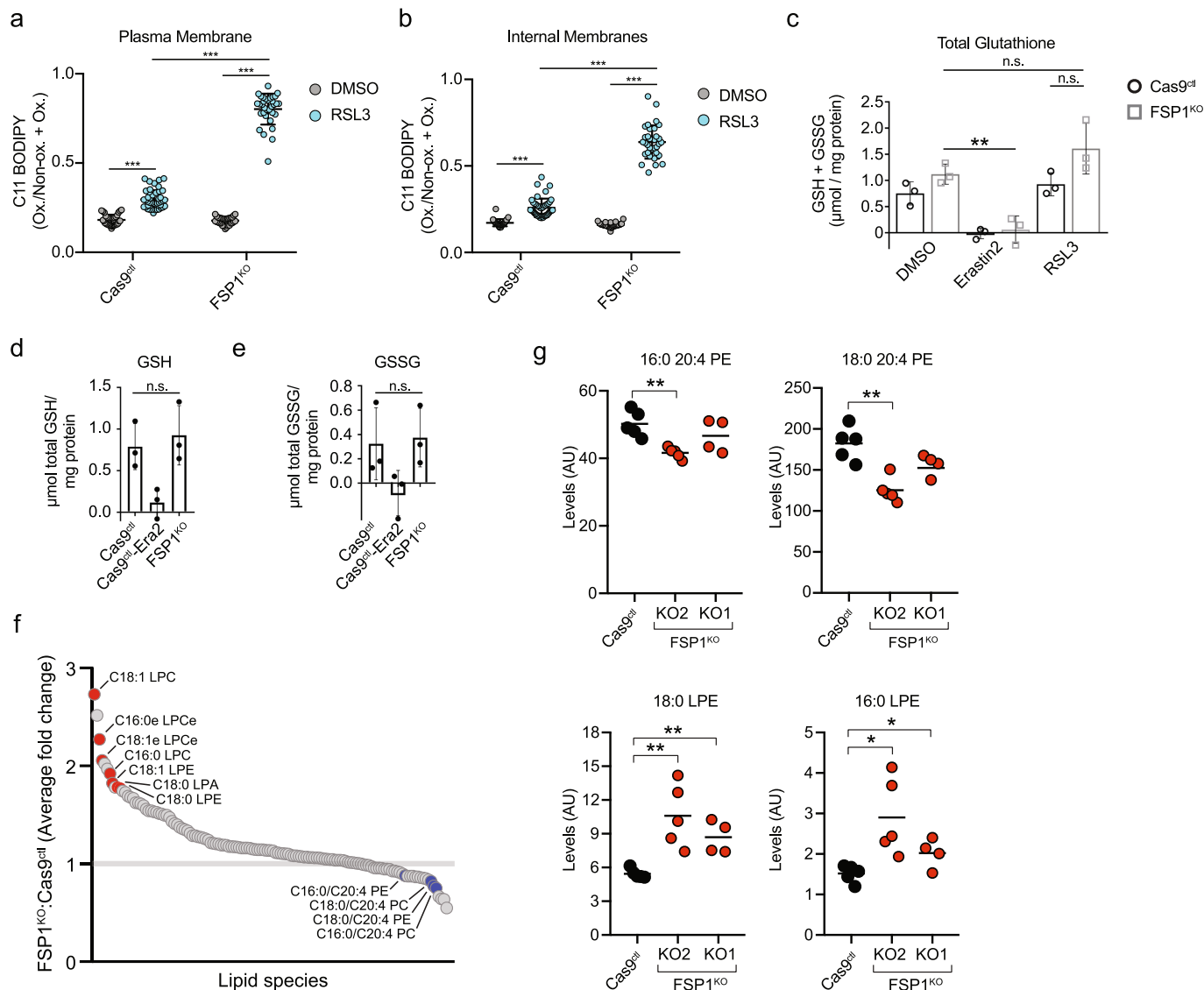
a, Western blot analysis of FSP1^{KO} cells induced for 48 h with doxycycline to express the indicated proteins. **b**, Live-cell microscopy of cells that express the indicated FSP1(G2A)-GFP constructs, incubated with 100 nM Mitotracker Orange to label mitochondria, 1 μ M BODIPY 558/568 C12 to label lipid droplets or 5 μ g ml⁻¹ Cell Mask to label the plasma membrane. To label the endoplasmic reticulum, cells were transiently transfected with BFP-Sec61 48 h before

imaging. Images are representative of at least $n = 10$ imaged cells. Line intensity plots show colocalization between FSP1 and organelle markers. Scale bar, 10 μ m. **c**, Plasma-membrane subdomains from FSP1^{KO} cells that express inducible LYN11-FSP1(G2A)-GFP were enriched by OptiPrep gradient centrifugation. The densitometry plot indicates the distribution of overexpressed and endogenous proteins. Panels are representative of two biological replicates except for **c**, which shows a single experiment.



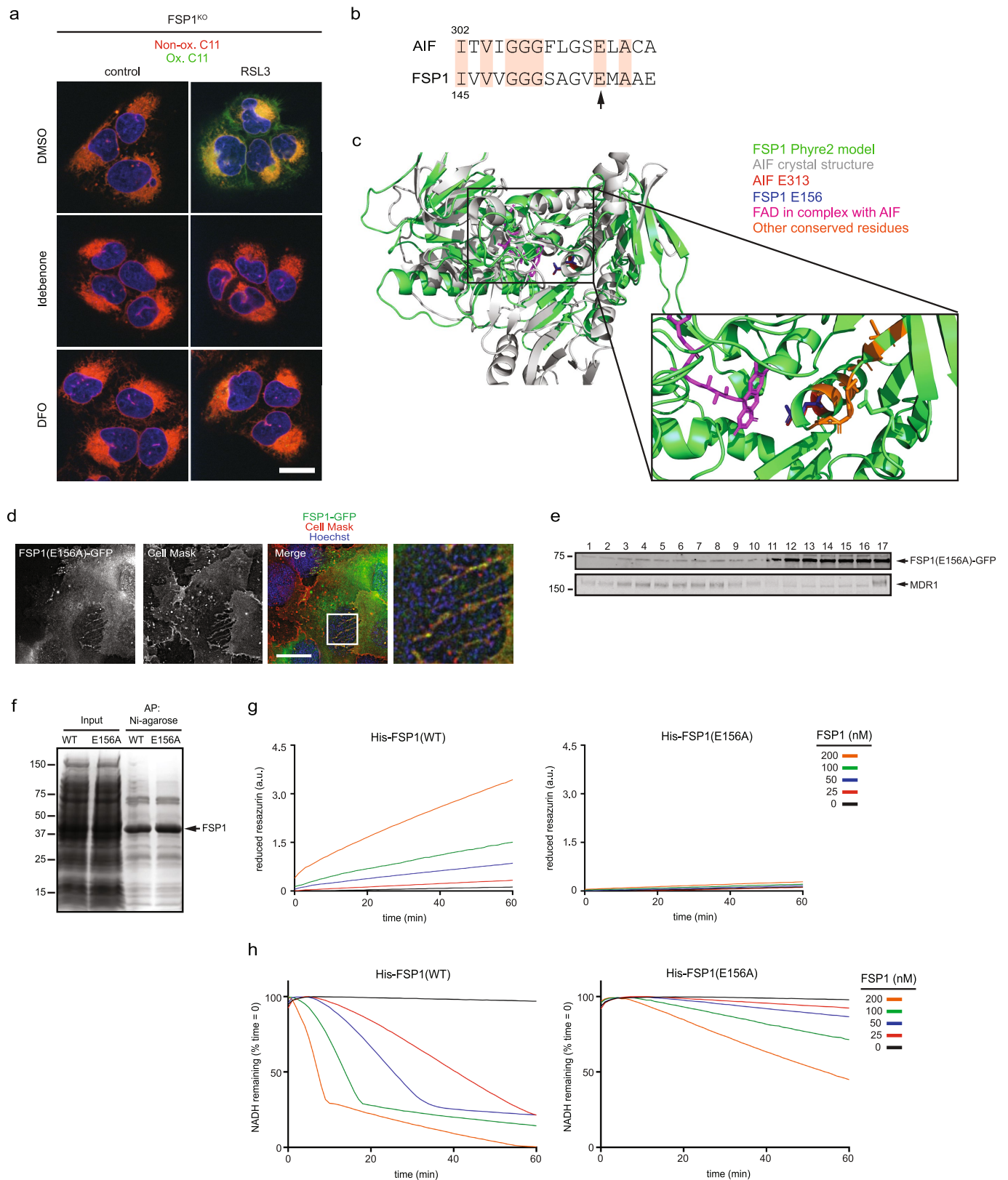
Extended Data Fig. 5 | Lipid droplets are not required for inhibition of ferroptosis by FSP1. a, Control cells were treated with inhibitors of DGAT1 (20 μM T863) and DGAT2 (10 μM PF-06424439) for 48 h, stained with 1 μM BODIPY 493/503 and imaged by fluorescence microscopy. The image is representative of $n = 50$ imaged fields. Scale bar, 10 μm . **b,** The size and number

of lipid droplets were quantified from cells ($n > 5,000$) in **a**. **c,** Dose response of RSL3-induced cell death of control cells pretreated for 48 h with 20 μM T863 and 10 μM PF-06424439 before addition of RSL3. Each data point is the average of three technical replicates. All panels are representative of two biological replicates.



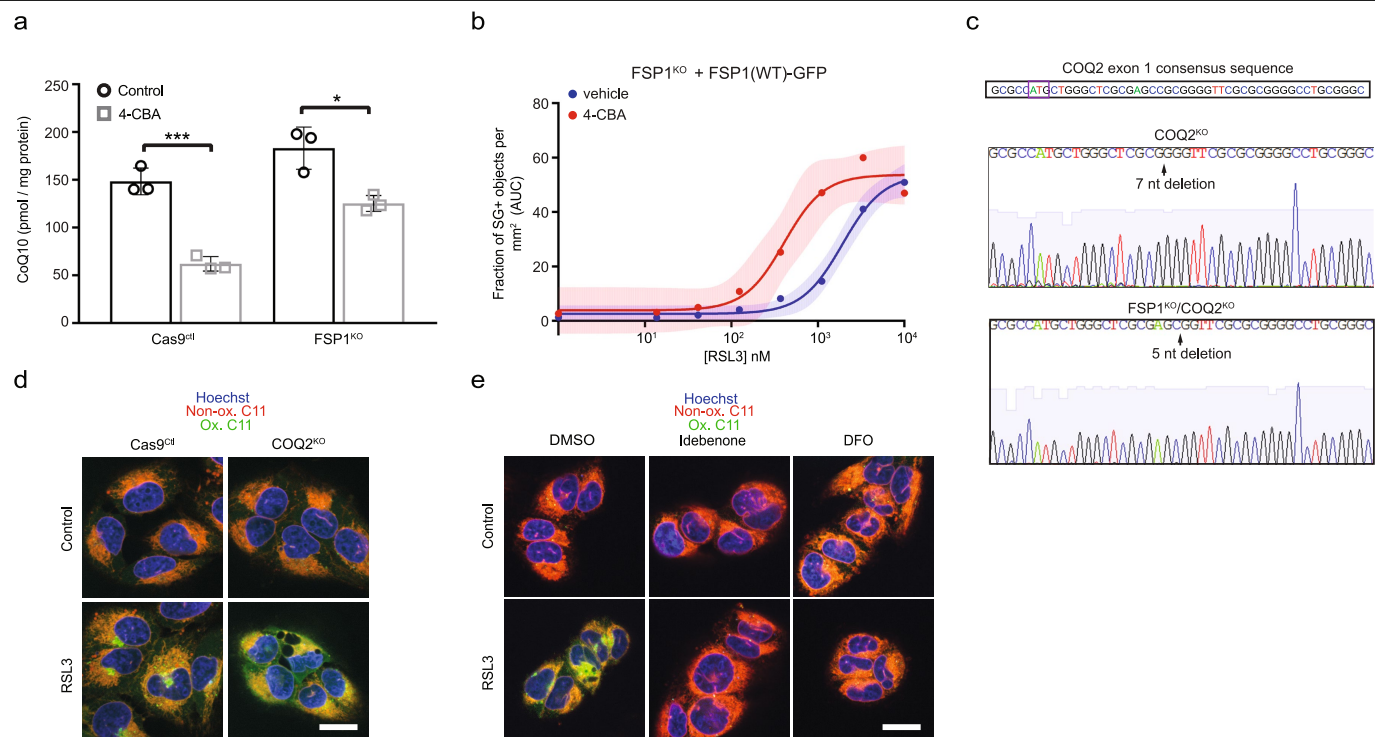
Extended Data Fig. 6 | Analysis of lipid peroxidation, glutathione and lipid levels in *FSP1*^{KO} cells. **a, b, Ratio of oxidized-to-total BODIPY 581/591 C11 from images in Fig. 3a, at the plasma membrane (**a**) or at internal membranes (**b**). Each data point represents an individual cell quantified in one of two biological replicates. For **a**, Cas9^{ctrl} DMSO, *n* = 34; Cas9^{ctrl} RSL3, *n* = 45; FSP1^{KO} DMSO, *n* = 30; FSP1^{KO} RSL3, *n* = 33; ****P* < 0.001 by one-way ANOVA. For **b**, Cas9^{ctrl} DMSO, *n* = 33; Cas9^{ctrl} RSL3, *n* = 45; FSP1^{KO} DMSO, *n* = 30; FSP1^{KO} RSL3, *n* = 33; ****P* < 0.001 by one-way ANOVA. Error bars show mean ± s.d. **c**, Total intracellular glutathione (GSH + GSSG) levels in control and FSP1^{KO} were determined following treatment with 250 nM RSL3 or 1 μM Erastin2. The graph shows mean ± s.d. of three biological replicates. n.s., FSP1^{KO} DMSO versus RSL3, *P* = 0.7278; n.s., FSP1^{KO} RSL3 versus Cas9^{ctrl} RSL3, *P* = 0.1522, ***P* = 0.0072 by one-way ANOVA.**

d, e, GSH and GSSG levels in control and FSP1^{KO} cells were measured. Where indicated, cells were treated with 1 μM Erastin2. The graph shows mean ± s.d. of three biological replicates. n.s., GSH *P* = 0.6269; n.s., GSSG *P* = 0.8284 by two-tailed *t*-test. **f**, The plot shows the average of the fold change in lipids measured in two FSP1^{KO} cell lines generated using *FSP1* sgRNA no. 1 and *FSP1* sgRNA no. 2 (labelled KO1 and KO2, respectively), relative to control cells. Cas9^{ctrl}, *n* = 5; KO1, *n* = 4; KO2, *n* = 5 biological replicates (Supplementary Table 3). **g**, Levels of select lipid species in biological replicates of control and FSP1^{KO} cells measured in **f**. The average values are indicated. 16:0 20:4 PE, ***P* = 0.0017; 18:0 20:4 PE, ***P* = 0.0011; 18:0 LPE, KO2 ***P* = 0.0036, KO1 ***P* = 0.0019; 16:0 LPE, KO2 **P* = 0.0133 and KO1 **P* = 0.0335 by two-tailed *t*-test.



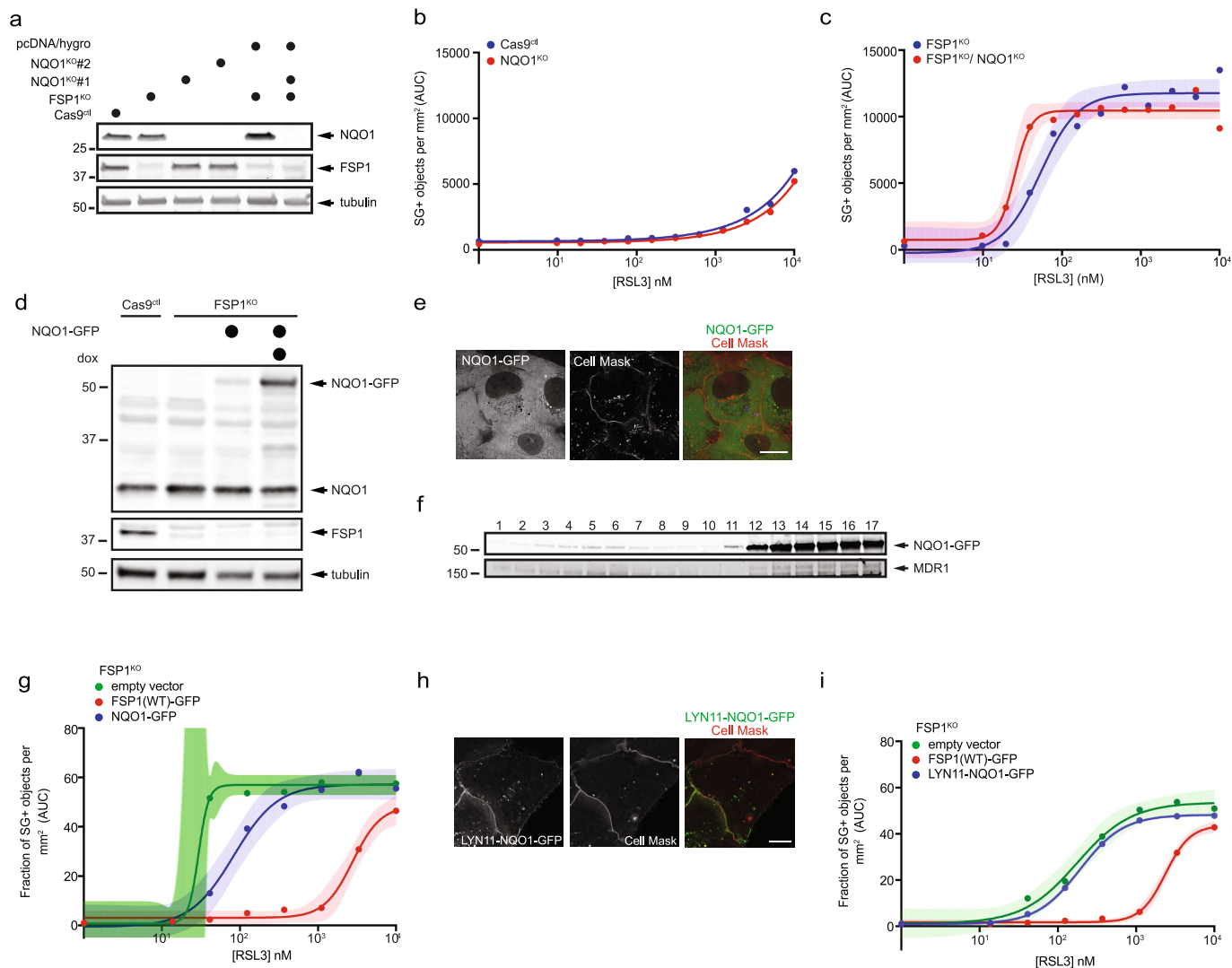
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Analysis of the FSP1 oxidoreductase mutant. **a**, FSP1^{KO} cells were treated with 250 nM RSL3 and 10 μ M idebenone or 50 μ M DFO for 75 min, labelled with BODIPY 581/591 C11 and fixed before imaging. Images are representative of at least $n = 10$ cells imaged for each treatment condition. Scale bar, 20 μ m. **b**, Sequence alignment showing residues conserved between AIF and FSP1. The arrow points to E313 in AIF (aligns to E156 in FSP1) that functions in binding to flavin adenine dinucleotide. **c**, Structural alignment between the crystal structure of mouse AIF (RCSB Protein Data Bank code (PDB) 1GV4) and the Phyre2-generated model of FSP1. **d**, Live-cell microscopy of FSP1^{KO} cells expressing inducible FSP1(E156A)–GFP labelled with 5 μ g ml⁻¹ Cell Mask. The image is representative of at least $n = 10$ imaged cells. Scale bar, 10 μ m. **e**, Plasma-membrane subdomains from FSP1^{KO} cells that express FSP1(E156A)–GFP were enriched by OptiPrep gradient centrifugation. **f**, SDS–PAGE and Coomassie brilliant blue stain of recombinant His–FSP1(WT) and His–FSP1(E156A) purified with Ni–NTA agarose beads. **g**, Reduction of resazurin by recombinant FSP1 in the presence of NADH. **h**, Oxidation of NADH by recombinant FSP1 in the presence of coenzyme Q₁. Panels **g** and **h** are representative of two biological replicates, and **e** shows a single experiment.



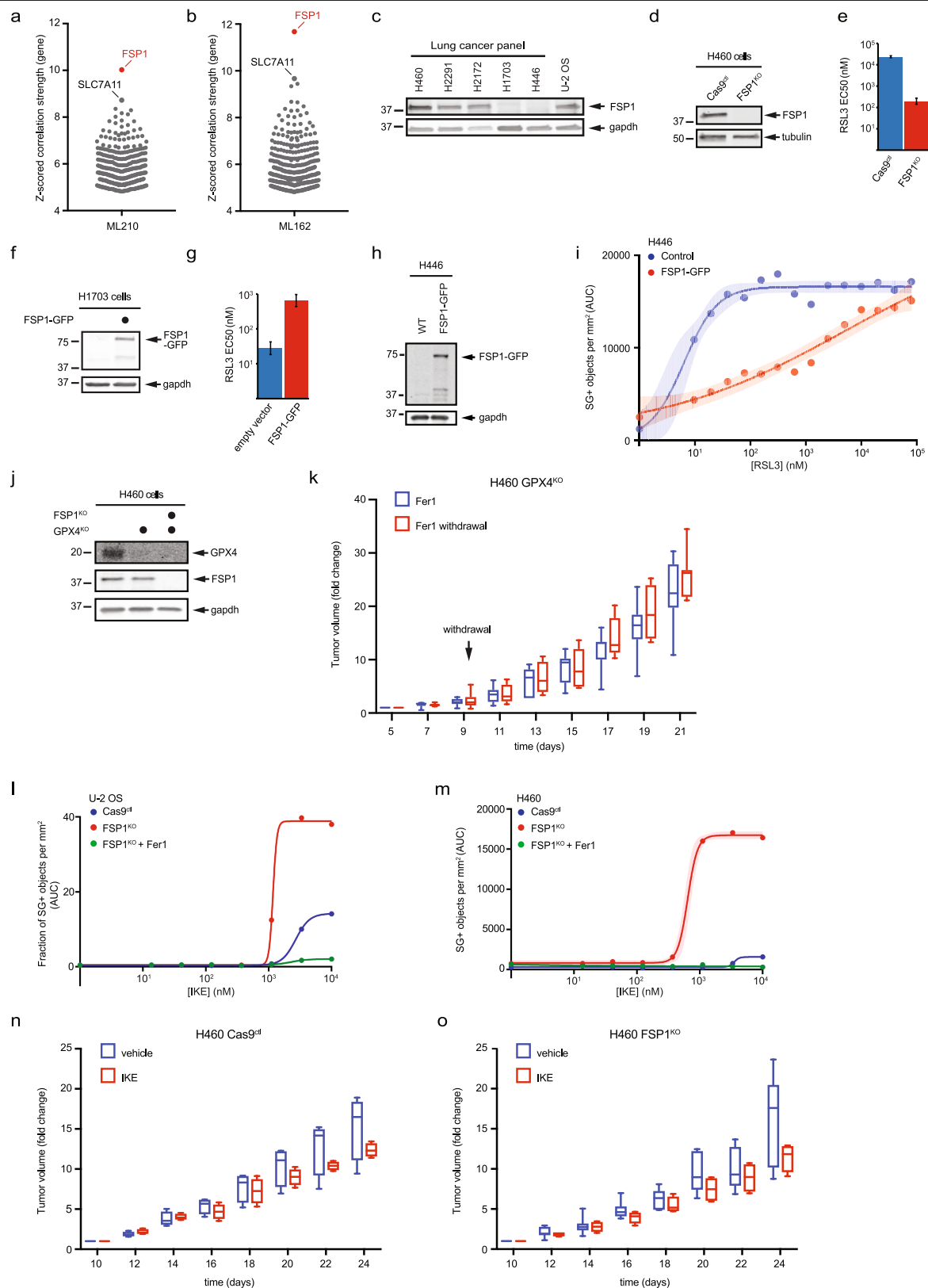
Extended Data Fig. 8 | Lipid peroxidation in CoQ-depleted cells. a, Total CoQ levels in control and FSP1^{KO} cells treated for 48 h with 3 mM 4-CBA. The graph shows mean \pm s.d. of three biological replicates. *** P =0.0007, * P =0.0132 by two-tailed t -test. **b**, Dose response of RSL3-induced death of inducible FSP1-GFP cells pretreated for 48 h with 3 mM 4-CBA and doxycycline before addition of RSL3. Shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. The panel is representative of two biological replicates. **c**, Genomic sequencing of the

COQ2 gene in COQ2^{KO} and FSP1^{KO} COQ2^{KO} cells. The ATG start codon is boxed in the *COQ2* consensus sequence. **d**, Control and COQ2^{KO} cells treated with 250 nM RSL3 for 3 h were labelled with BODIPY 581/591 C11 and fixed before imaging. **e**, COQ2^{KO} cells were treated with 250 nM RSL3 and 10 μ M idebenone or 50 μ M DFO for 3 h, labelled with BODIPY 581/591 C11 and fixed before imaging. In panels **d**, **e**, images are representative of at least n =10 cells imaged for each treatment condition. Scale bars, 20 μ m.



Extended Data Fig. 9 | Role of NQO1 in ferroptosis resistance. **a**, Western blot analysis of lysates from NQO1^{KO} and NQO1^{KO} FSP1^{KO} cells. **b**, Dose response of RSL3-induced death of control and NQO1^{KO} cells. **c**, Dose response of RSL3-induced death of FSP1^{KO} and NQO1^{KO} FSP1^{KO} cells. Cells in **b** and **c** were generated using *NQO1* sgRNA 1. **d**, Western blot analysis of lysates of FSP1^{KO} cells that express doxycycline-inducible NQO1-GFP. **e**, Live-cell microscopy of inducible NQO1-GFP cells labelled with 5 $\mu\text{g ml}^{-1}$ Cell Mask. **f**, Plasma-membrane subdomains from FSP1^{KO} cells that express NQO1-GFP were enriched by OptiPrep gradient centrifugation. **g**, Dose response of RSL3-induced death of

FSP1^{KO} cells expressing the indicated inducible constructs. **h**, Live-cell microscopy of FSP1^{KO} cells that express inducible LYN11-NQO1-GFP labelled with 5 $\mu\text{g ml}^{-1}$ Cell Mask. **i**, Dose response of RSL3-induced death of FSP1^{KO} cells that express the indicated inducible constructs. For panels **b**, **c**, **g**, **i**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates. Panels are representative of two biological replicates except for **f** and **i**, which show the results of single experiments. In **e** and **h**, the images are representative of at least $n = 10$ imaged cells. Scale bars, 10 μm .



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The role of FSP1 in cancer. **a, b**, A high level of expression of FSP1 is correlated with resistance to the GPX4 inhibitors ML210 (**a**) and ML162 (**b**) in non-haematopoietic cancer cells. Plotted data were mined from the CTRP database that contains correlation coefficients between gene expression and drug sensitivity for 907 cancer cell lines treated with 545 compounds. Plotted values are z-scored Pearson's correlation coefficients. **c**, Western blot of FSP1 expression in a panel of lung cancer lines. **d**, Western blot of lysates from control and FSP1^{KO} H460 cells. **e**, EC₅₀ RSL3 dose for the indicated H460 cell lines was calculated from the results in Fig. 1d. Bars indicate 95% confidence intervals. **f**, Western blot of lysates from control and H1703 cells. **g**, EC₅₀ RSL3 dose for the indicated H1703 cell lines was calculated from the results in Fig. 1e. Bars indicate 95% confidence intervals. **h**, Western blot analysis of H446 cells that express doxycycline-inducible FSP1-GFP. **i**, Dose response of RSL3-induced death of control and FSP1-GFP H446 cells. **j**, Western blot analysis of GPX4^{KO} and GPX4^{KO} FSP1^{KO} H460 cells. **k**, GPX4^{KO} H460 tumour xenograft cells were initiated in immune-deficient SCID mice (*n* = 16). Following 5 days of daily Fer1 injections (2 mg kg⁻¹ body weight) to allow lines to

develop tumours, 1 set of mice (*n* = 8) continued to receive daily Fer1 injections and a second set (*n* = 8) received vehicle injections for the remaining 17 days. The distribution of fold changes in sizes of individual tumours during the treatment is shown. GPX4^{KO} (-) Fer1, *n* = 7; GPX4^{KO} (+) Fer1, *n* = 7. **l**, Dose response of IKE-induced death of control and FSP1^{KO} U-2 OS cells. **m**, Dose response of IKE-induced death of control and FSP1^{KO} H460 cells. **n, o**, Control (**n**) and FSP1^{KO} (**o**) H460 tumour xenografts were initiated in immune-deficient SCID mice (*n* = 16). After 10 days, each group of mice (*n* = 8) was injected daily with IKE or vehicle (40 mg kg⁻¹ body weight). The distribution of fold changes in sizes of individual tumours during the treatment is shown. Cas9^{ctrl} (-) IKE, *n* = 4; Cas9^{ctrl} (+) IKE, *n* = 4; FSP1^{KO} (-) IKE, *n* = 7; FSP1^{KO} (+) IKE, *n* = 4. In **k, n, o**, box plots show median, 25th and 75th percentiles, minima and maxima of the distributions. Panels are representative of two biological replicates except **l, m**, which show the results of single experiments. In **i, l, m**, shading indicates 95% confidence intervals for the fitted curves and each data point is the average of three technical replicates.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	SoftWoRx V6.5.2 (GE Life Sciences), PHYRE2 Protein Fold Recognition Software V2.0 (www.sbg.bio.ac.uk/phyre2/), BD FACSDiva V6.2 (bdbiosciences.com), SoftMax Pro V6.3 (moleculardevices.com), iQ3 live cell imaging software (Andor Technology), EZChrom Elite V3.2.0 (Agilent), Image Lab V6.0.1 (Bio-Rad Laboratories, Inc.)
Data analysis	casTLE statistical framework V1.0 (bitbucket.org/dmorgens/castle/), ImageJ V1.8.0 (imagej.nih.gov/ij/), MATLAB R2016b (mathworks.com), droplet detection and quantification software for MATLAB (Olzmann lab, www.dropletproteome.org), Prism V7 (GraphPad), Zoom Image Analysis Software 2016B (Essen Bioscience), BowTie 2 V2.3.4.3 (bowtie-bio.sourceforge.net/bowtie2/index.shtml), FlowJo V10 (TreeStar) (flowjo.com)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data that support the conclusions in this manuscript are available from the corresponding author upon request. Raw data for figure 1 can be accessed in Supplementary Table 1. Raw data for figure 3 can be accessed in Supplementary Table 3. Raw data for figure 4 can be accessed in Supplementary Table 4 and are publicly available from the CTRP and CCLE databases (portals.broadinstitute.org).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical tests were used to calculate sample size. In cases where statistics were derived, sample size was n=3 or more independent biological replicates. For measurement of lipid levels using mass spectrometry and measurement of reduced COQ levels using HPLC, sample sizes were n=5 and n=6 biological replicates, respectively, to account for expected variability due to sample preparation and noise from the instruments. For mouse xenograft experiments, sample size was n=8 for each treatment group to account for differences in tumor formation and growth, and to ensure recovery of a sufficient quantity of mice with successful xenografts of approved size at each time point of the study.
Data exclusions	These criteria were established prior to performing the xenograft studies. In the ferrostatin-1 withdrawal experiments, animals not included in the analysis included mice that were sacrificed early due to sickness (n = 1 of GPX4KO (+) Fer1) and mice whose tumors were determined to be outliers according to the Grubbs statistical test using Prism (Graphpad) software (n = 1 of GPX4KO (-) Fer1 and n = 1 of GPX4KO/FSP1KO (-) Fer1). For the IKE injection experiments, animals not included in the analysis included mice that were sacrificed early due to development of exceedingly large tumors (n = 1 of Cas9 ctl (+) IKE, n = 3 of Cas9 ctl (-) IKE, n = 3 of FSP1KO (+) IKE and n = 1 of FSP1KO (-) IKE), mice in which tumors failed to initiate (n = 2 of Cas9 ctl (+) IKE), and mice whose tumors were determined to be statistical outliers according to the Grubbs test (n = 1 of Cas9 ctl (+) IKE, n = 1 of Cas9 ctl (-) IKE, n = 1 of FSP1KO (+) IKE).
Replication	All attempts at replication were successful. Figures, including western blots, dose response curves and enzymatic activity assay panels are representative of two biological replicates except for the following, which show single experiments: plasma membrane fractionations in Extended Data Fig. 4c, 7e and 9f and cell death curves in Extended Data Fig. 1d,e, 1k,j, 9i and 10l,m. Images are representative of at least n = 10 imaged cells.
Randomization	For the xenograft studies, following injection of H460 cells, the mice were randomly assigned into 2 treatment groups for the ferrostatin-1 withdrawal experiments and into 2 treatment groups for the IKE injection experiments.
Blinding	Blinding was not possible because the experiments were performed by a single researcher.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Anti-Plin2 (Abgent, cat. #AP5118c), anti-AIFM2 (Proteintech Group, Inc. cat. #20886-1-AP and Santa Cruz Biotechnology, clone B-6, cat. #sc-377120), anti- α -tubulin (Cell Signaling Technology, Inc., cat. #2144 and Santa Cruz Biotechnology, clone B-7, cat. #sc-5286), anti-GPX4 (Abcam, cat. #ab41787), anti-ACSL4 (Sigma-Aldrich, cat. #SAB-2701949), anti-GFP (Poteintech Group, Inc., clone 1E10H7, cat. #66002-1-1), anti-NQO1 (Proteintech Group, Inc., cat. #1145-1-AP), anti-GAPDH (EMD Millipore, cat. #mab374), anti-RAS (Cell Biolabs, Inc. cat. #STA-400), anti-MDR1 (Cell Signaling Technology, Inc., clone D3H1Q, cat. #12683S), anti-p21 (Cell Signaling Technology, Inc., clone 12D1, cat. #2947).
Validation	Anti-AIFM2, anti-GPX4, ant-ACSL4, and anti-NQO1 were validated using genetic knockout of the endogenous genes with Cas9 and one or more targeted sgRNAs. Anti-Plin2, anti- α -tubulin, anti-GAPDH, anti-RAS, anti-MDR1 and anti-p21 were validated un in human cells by the manufacturer.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	H460, H2291, H2172, H1703, H446 were purchased from ATCC (atcc.org). U-2 OS Flp-In cells were a gift from Dr. Daniel Durocher (Lunenfeld-Tenenbaum Research Institute).
Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	All cell lines are negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Cell lines used in the study are not flagged in the Register of Misidentified Cell Lines.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male, C.B17 SCID mice, 6 weeks of age
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	All animal experiments were conducted in accordance with the guidelines of the Institutional Animal Care and Use Committees (IACUC) of the University of California, Berkeley.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells grown in 6-cm plates were washed with PBS, trypsinized and centrifuged for 5 min at 500 x g. Cell pellets were resuspended in PBS containing 5% FBS and 5 μ M CellEvent™ Caspase-3/7 Green Detection Reagent, and were incubated for 30 min at 37°C prior to analysis.
Instrument	LSRFortessa (Becton-Dickinson)
Software	Data was collected using BD FACSDiva V6.2 (Becton-Dickinson) and analyzed using FlowJo V10 (TreeStar).
Cell population abundance	Cells were not sorted during the procedure.
Gating strategy	Apoptotic cells were gated using the same low FSC threshold (FSC- gate) across all samples and the FITC signals of the gated populations were determined.

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

FSP1 is a glutathione-independent ferroptosis suppressor

<https://doi.org/10.1038/s41586-019-1707-0>

Received: 25 February 2019

Accepted: 9 October 2019

Published online: 21 October 2019

Sebastian Doll^{1,13}, Florencio Porto Freitas^{2,13}, Ron Shah³, Maceler Aldrovandi^{1,4}, Milene Costa da Silva¹, Irina Ingold¹, Andrea Goya Grocin⁵, Thamara Nishida Xavier da Silva², Elena Panzilius⁶, Christina H. Scheel^{6,7}, André Mourão⁸, Katalin Buday¹, Mami Sato¹, Jonas Wanninger¹, Thibaut Vignane¹, Vaishnavi Mohana¹, Markus Rehberg⁹, Andrew Flatley¹⁰, Aloys Schepers¹⁰, Andreas Kurz¹¹, Daniel White⁴, Markus Sauer¹¹, Michael Sattler⁸, Edward William Tate⁵, Werner Schmitz¹², Almut Schulze¹², Valerie O'Donnell⁴, Bettina Proneth¹, Grzegorz M. Popowicz⁸, Derek A. Pratt³, José Pedro Friedmann Angeli^{2*} & Marcus Conrad^{1*}

Ferroptosis is an iron-dependent form of necrotic cell death marked by oxidative damage to phospholipids^{1,2}. To date, ferroptosis has been thought to be controlled only by the phospholipid hydroperoxide-reducing enzyme glutathione peroxidase 4 (GPX4)^{3,4} and radical-trapping antioxidants^{5,6}. However, elucidation of the factors that underlie the sensitivity of a given cell type to ferroptosis⁷ is crucial to understand the pathophysiological role of ferroptosis and how it may be exploited for the treatment of cancer. Although metabolic constraints⁸ and phospholipid composition^{9,10} contribute to ferroptosis sensitivity, no cell-autonomous mechanisms have been identified that account for the resistance of cells to ferroptosis. Here we used an expression cloning approach to identify genes in human cancer cells that are able to complement the loss of GPX4. We found that the flavoprotein apoptosis-inducing factor mitochondria-associated 2 (*AIFM2*) is a previously unrecognized anti-ferroptotic gene. *AIFM2*, which we renamed ferroptosis suppressor protein 1 (FSP1) and which was initially described as a pro-apoptotic gene¹¹, confers protection against ferroptosis elicited by *GPX4* deletion. We further demonstrate that the suppression of ferroptosis by FSP1 is mediated by ubiquinone (also known as coenzyme Q₁₀, CoQ₁₀): the reduced form, ubiquinol, traps lipid peroxyl radicals that mediate lipid peroxidation, whereas FSP1 catalyses the regeneration of CoQ₁₀ using NAD(P)H. Pharmacological targeting of FSP1 strongly synergizes with GPX4 inhibitors to trigger ferroptosis in a number of cancer entities. In conclusion, the FSP1–CoQ₁₀–NAD(P)H pathway exists as a stand-alone parallel system, which co-operates with GPX4 and glutathione to suppress phospholipid peroxidation and ferroptosis.

Ferroptosis is controlled by the selenoenzyme GPX4^{3,4,12}. With the recognition that targeting ferroptosis may help to eradicate therapy-resistant tumours in patients^{13–15}, there is mounting interest in understanding the mechanisms that underpin the sensitivity of cells to ferroptosis¹⁶. Although acyl-CoA synthetase long chain family member 4 (*ACSL4*) was identified as a pro-ferroptotic gene, the expression of which determines ferroptosis sensitivity^{9,10}, inhibition of GPX4 fails to trigger ferroptosis in some cancer cell lines regardless of *ACSL4* expression, suggesting that there are alternative resistance mechanisms.

Genetic suppressor screen uncovers FSP1

To uncover these factors, we generated a cDNA expression library derived from the MCF7 ferroptosis-resistant cell line^{9,10} (Extended Data Fig. 1a), and screened for genes complementing loss of *GPX4* (Fig. 1a). Sequencing of 14 single-cell clones identified 7 clones that express either *GPX4* or *AIFM2* (Extended Data Fig. 1b). *AIFM2* is a flavoprotein that has originally been described as a p53-responsive gene¹⁷ and claimed to induce apoptosis based on sequence similarity to another initially postulated pro-apoptotic gene,

¹Institute of Developmental Genetics, Helmholtz Zentrum München, Neuherberg, Germany. ²Rudolf Virchow Center for Experimental Biomedicine, University of Würzburg, Würzburg, Germany.

³Department of Chemistry & Biomolecular Sciences, University of Ottawa, Ottawa, ON, Canada. ⁴Systems Immunity Research Institute, School of Medicine, Cardiff University, Cardiff, UK.

⁵Molecular Sciences Research Hub, Department of Chemistry, Imperial College London, London, UK. ⁶Institute of Stem Cell Biology, Helmholtz Zentrum München, Neuherberg, Germany.

⁷Clinic for Dermatology, St Josef Hospital Bochum, University of Bochum, Bochum, Germany. ⁸Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany. ⁹Institute of Lung Biology and Disease, Helmholtz Zentrum München, Neuherberg, Germany. ¹⁰Monoclonal Antibody Core Facility, Helmholtz Zentrum München, Neuherberg, Germany. ¹¹Department of

Biotechnology & Biophysics, Biocenter, University of Würzburg, Würzburg, Germany. ¹²Department of Biochemistry and Molecular Biology, Theodor Boveri Institute, Biocenter, University of Würzburg, Würzburg, Germany. ¹³These authors contributed equally: Sebastian Doll, Florencio Porto Freitas. *e-mail: pedro.angeli@virchow.uni-wuerzburg.de;

marcus.conrad@helmholtz-muenchen.de

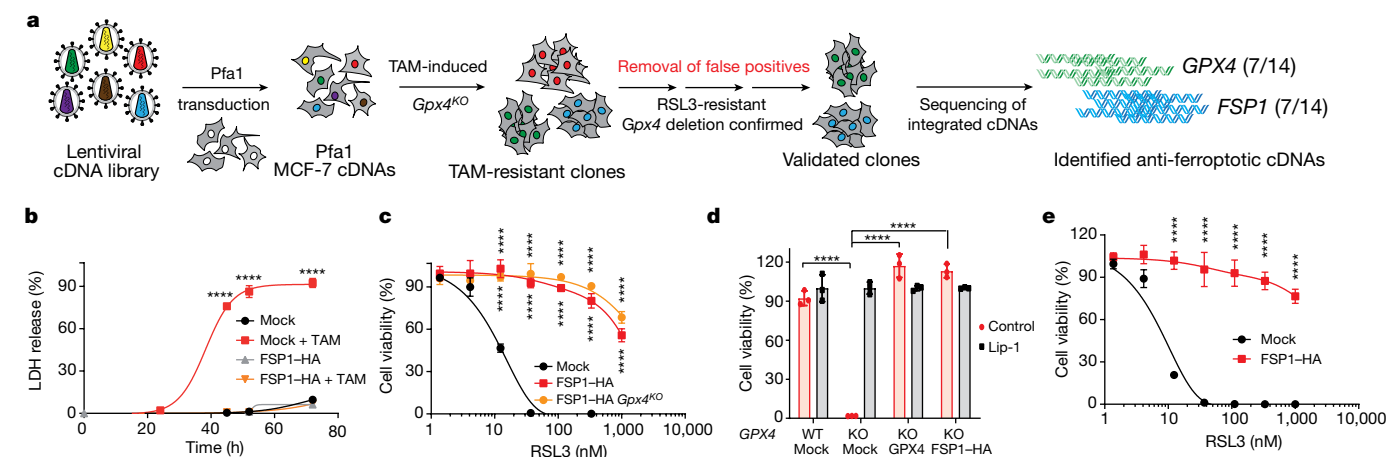


Fig. 1 | Identification and validation of FSP1 as a ferroptosis suppressor.

a, Schematic of the identification of FSP1 as a ferroptosis suppressor, using double selection with 4-hydroxytamoxifen (TAM)-induced *Gpx4*-knockout (KO) followed by RSL3-mediated elimination of false-positive cell clones. Surviving single-cell clones were analysed by Sanger sequencing. **b**, Cell death induced by TAM was measured by lactate dehydrogenase (LDH) release of Pfa1 cells stably expressing an empty vector (mock) and FSP1 tagged with haemagglutinin (FSP1-HA) using supernatants collected at the indicated time points in a 96-well plate. **c–e**, Dose-dependent toxicity of RSL3-induced cell

death of Pfa1 *Gpx4* wild-type (WT) or *Gpx4*-knockout for cells expressing mock or FSP1-HA (**c**), *GPX4* wild-type and *GPX4*-knockout HT1080 cells overexpressing mock, human *GPX4*-FSH (Flag-Strep-His-tag) or FSP1-HA treated with or without 200 nM liproxstatin-1 (Lip-1; **d**) and HT1080 cells expressing mock or FSP1-HA (**e**). Cell viability was assessed 24 h (**c**, **e**) or 72 h (**d**) thereafter using Aquabluor. Data are the mean \pm s.d. of $n = 3$ wells of a 96-well plate from one representative of two (**b**) or three (**c–e**) independent experiments; **** $P < 0.0001$; two-way analysis of variance (ANOVA).

apoptosis-inducing factor mitochondria-associated 1 (AIFM1)¹¹. To avoid further confusion, we therefore recommend that in the future AIFM2 is referred to as ferroptosis suppressor protein 1 (FSP1)¹⁸. For validation, we stably expressed FSP1 in mouse Pfa1¹⁹ and in human fibrosarcoma HT1080 cells (Extended Data Fig. 1c, d). FSP1-overexpressing cells were robustly protected against pharmacological and genetic inducers of ferroptosis¹ and proliferated indefinitely (Fig. 1b–e, Extended Data Fig. 1e–i and Supplementary Video 1). To our knowledge, this is the first enzymatic system that complements loss of *GPX4*¹⁹.

The anti-ferroptotic function of FSP1 was found to be independent of cellular glutathione levels, GPX4 activity, ACSL4 expression and oxidizable fatty acid content (Extended Data Figs. 1c, d, j, k, 2), showing that FSP1 does not interfere with canonical ferroptosis mechanisms. Moreover, the protection against cell death conferred by FSP1 was specific to ferroptosis-inducing agents; FSP1 did not protect against cell death caused by cytotoxic compounds and/or pro-apoptotic conditions. Moreover, p53 status did not affect FSP1 expression (Extended Data Fig. 3a–e). In contrast to FSP1, overexpression of AIFM1 failed to suppress ferroptosis (Extended Data Fig. 3f, g).

N-myristoylation enables ferroptosis resistance

Our early efforts revealed that N-terminal tagging of FSP1 abolished its anti-ferroptotic activity. Indeed, the N terminus of FSP1 contains a canonical myristoylation motif²⁰, which presumably facilitates its association with lipid bilayers²¹. Expression of wild-type FSP1 and a mutant form of FSP1 that lacks the predicted myristoylation site (G2A) in Pfa1 and HT1080 cells (Fig. 2a), as well as FSP1 tagging with an alkyne-functionalized myristic acid mimetic (YnMyr) enabled the specific enrichment of only wild-type FSP1. This enrichment was abolished either in FSP1(G2A) mutants or after treatment with the pan-N-myristoyl transferase inhibitor IMP-1088²² (Fig. 2b). Myristoylation of FSP1 appears to be essential for its anti-ferroptotic activity as FSP1(G2A)- and wild-type FSP1-expressing cells treated with IMP-1088 showed abrogated ferroptosis resistance (Fig. 2c, d and Extended Data Fig. 3h, i). We therefore assessed the subcellular

distribution of both wild-type FSP1 and FSP1(G2A) using C-terminally tagged fusion proteins. Although FSP1–GFP localized to an unspecified perinuclear membrane compartment, it also partially overlapped with endoplasmic reticulum and Golgi markers (Fig. 2e and Extended Data Fig. 4a). By contrast, FSP1(G2A)–GFP was distributed throughout the cell, suggesting that ferroptosis is perhaps driven in a specific subcellular region. A more in-depth investigation of the subcellular localization of FSP1 is provided in a companion study¹⁸ that reveals a notable role of plasma membrane-targeted FSP1 in the suppression of ferroptosis.

FSP1 prevents lipid peroxidation

As ferroptosis is driven by phospholipid peroxidation (pLPO), we stained Pfa1 cells with BODIPY 581/591 C11 and found that FSP1 overexpression blunted lipid peroxidation induced by (1S, 3R)-RSL3 (RSL3; Fig. 3a). Moreover, specific pLPO products were markedly lower in *Gpx4*-knockout FSP1-overexpressing cells (Fig. 3b and Extended Data Fig. 4b). As members of the AIF family have been shown to possess NADH:ubiquinone oxidoreductase activity²³, we hypothesized that FSP1 suppresses pLPO by regenerating lipophilic radical-trapping antioxidants using NAD(P)H. The reduced form of coenzyme Q₁₀ (CoQ₁₀-H₂) has been reported to be a good radical-trapping antioxidant in phospholipids and lipoproteins²⁴, yet is considered to be of limited importance outside mitochondria, as an efficient recycling mechanism has not been described. To investigate a possible link between FSP1 and CoQ₁₀-H₂, we generated CoQ₁₀-deficient HT1080 cells by deleting 4-hydroxybenzoate polyprenyltransferase (COQ2), the enzyme that catalyses the first step in CoQ₁₀ biosynthesis (Fig. 3c). CoQ₁₀-deprived cells proliferated normally when supplemented with uridine, CoQ₁₀ or decyl-ubiquinone (Extended Data Fig. 4c). Notably, whereas FSP1–GFP overexpression in parental HT1080 cells suppressed ferroptosis, it failed to do so in *Coq2*-knockout cells (Fig. 3d and Extended Data Fig. 4d). Consistent with previous data that have shown that purified FSP1 reduces ubiquinone analogues of variable chain lengths²³, heterologously expressed FSP1 (Extended Data Fig. 4e) catalysed the reduction of an ubiquinone analogue with an appended coumarin fluorophore. This enabled

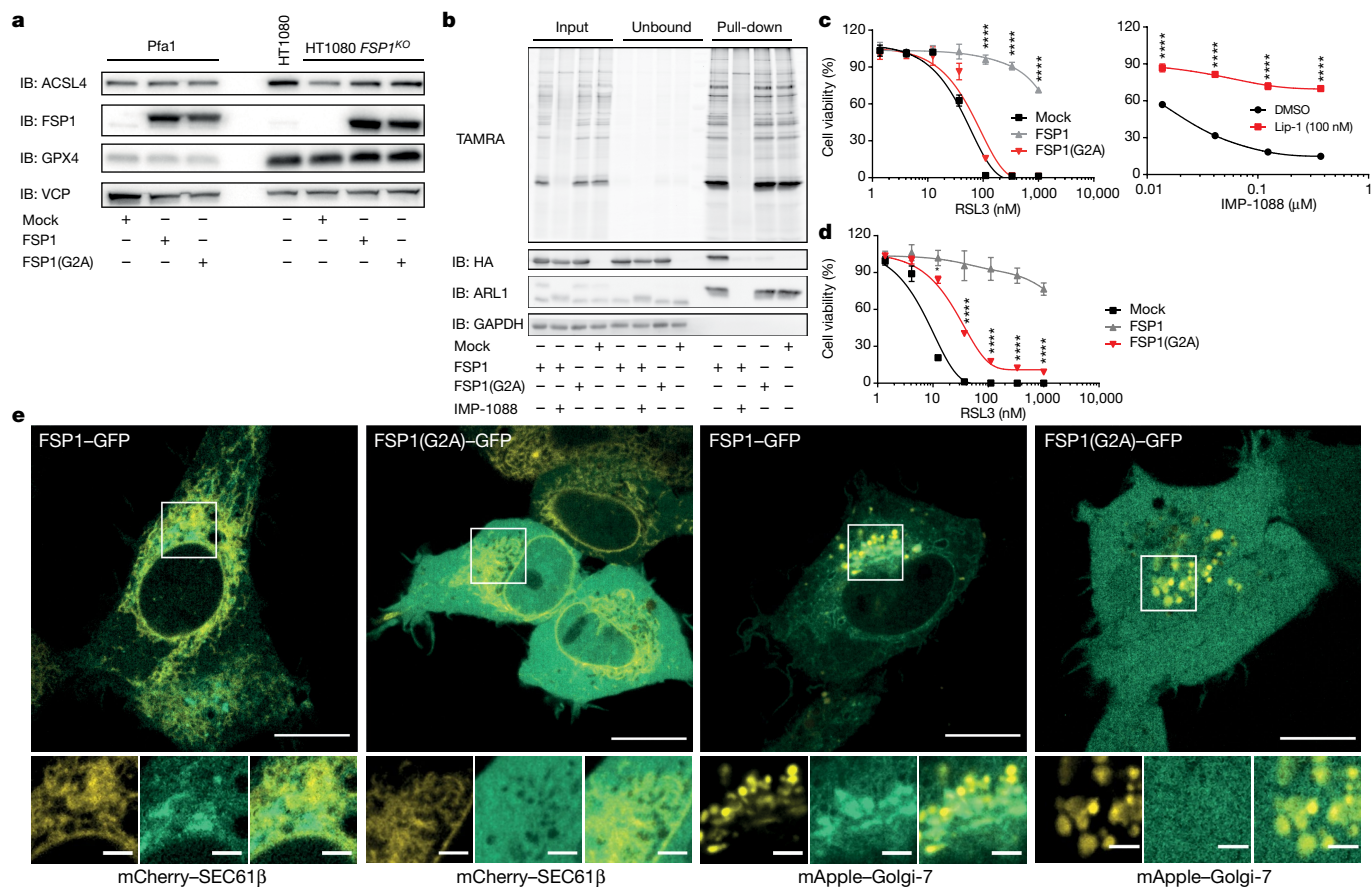


Fig. 2 | N-Myristoylation of FSP1 is important for its anti-ferroptotic function. **a**, Immunoblotting (IB) analysis of ACSL4, FSP1, GPX4 and valosin-containing protein (VCP) expression of Pfa1 cells stably expressing mock, FSP1-HA or FSP1(G2A)-HA (left), parental HT1080 cells and *FSP1*-knockout HT1080 cells stably expressing mock, FSP1-HA or FSP1(G2A)-HA (right). Immunoblot images are cropped from the chemiluminescence signal files. For gel source data showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1. **b**, Specific enrichment of myristoylated proteins using metabolic labelling with the YnMyr myristate analogue followed by click chemistry to AzTB (Pfa1 FSP1-HA, Pfa1 FSP1-HA and IMP-1088, Pfa1 FSP1(G2A)-HA, Pfa1 mock). TAMRA in-gel fluorescence showing labelling of myristoylated proteins. FSP1 was specifically enriched with YnMyr and the enrichment was prevented by the pan-myristoylation inhibitor IMP-1088 as well as by the FSP1(G2A) mutant, demonstrated by immunoblot analysis (HA antibody). Endogenously expressed ADP ribosylation factor-like GTPase 1 (ARL1), served as positive control and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as loading control. Immunoblot images are cropped from the chemiluminescence signal files.

For gel source data showing the Cy5 ladder and chemiluminescence signals separately, see Supplementary Fig. 1. **c**, Left, dose-dependent toxicity of RSL3 in Pfa1 cells stably expressing mock, FSP1-HA or FSP1(G2A)-HA. The FSP1(G2A) mutant failed to prevent RSL3-induced ferroptosis. Right, inhibition of myristoylation (IMP-1088) in FSP1-overexpressing *Gpx4*-knockout Pfa1 cells induced cell death in a dose-dependent manner, which was prevented by the ferroptosis inhibitor Lip-1. **d**, RSL3-induced cell death of *FSP1*-knockout HT1080 cells stably expressing mock, FSP1 or FSP1(G2A). Cell viability was assessed after 24 h using Aquabluor (**c**, **d**). Data are the mean \pm s.d. of $n = 4$ (**c**, left) or $n = 3$ (**c**, right; **d**) wells of a 96-well plate from one representative of three (**c**, **d**) independent experiments, **** $P < 0.0001$; two-way ANOVA. **e**, Enhanced resolution confocal microscopy of HT1080 cells (FSP1-GFP or FSP1(G2A)-GFP) overexpressing mCherry-SEC61β (endoplasmic reticulum localization) or mApple-Golgi-7 (Golgi localization). GFP is displayed in green; mCherry and mApple fluorescence are pseudo-coloured in yellow. Scale bars, 10 μm (top) and 2 μm (bottom, magnified images).

the determination of kinetic parameters for FSP1, which revealed a relatively low Michaelis constant ($K_m = 1.2 \times 10^{-5}$ M) and much higher maximum rate of the reaction ($V_{max} = 4.1 \times 10^{-7}$ M s⁻¹) compared to related oxidoreductases (for example, NQO1 ($K_m = 7.9 \times 10^{-7}$ M and $V_{max} = 6.1 \times 10^{-9}$ M s⁻¹)), as well as the expected inhibition of the substrate (Fig. 3e). Notably, we found that dehydroascorbate, oxidized glutathione and tert-butyl hydroperoxide did not act as substrates of FSP1 (Fig. 3f).

To further investigate our hypothesis that FSP1 suppresses pLPO by reducing CoQ₁₀, we carried out co-autoxidation experiments with egg phosphatidylcholine and STY-BODIPY²⁵ using a lipophilic alkoxyl radical generator (Extended Data Fig. 5a, b). We found that neither FSP1 alone or in combination with its reducing co-substrate, NAD(P)H, was able to suppress pLPO effectively (Extended Data Fig. 5c), whereas

addition of CoQ₁₀ substantially delayed the autoxidation of egg phosphatidylcholine in a dose-dependent manner (Extended Data Fig. 5d, e). These results suggest that, through FSP1, CoQ₁₀ helps to shuttle reducing equivalents from NAD(P)H into the lipid bilayer to inhibit propagation of lipid peroxidation. NQO1 was unable to serve in the same capacity as FSP1 in these assays (Extended Data Fig. 5f, g). As CoQ₁₀ is readily autoxidized and has poor dynamics within the lipid bilayer²⁶, we wondered whether α-tocopherol may also contribute to the protection against ferroptosis observed by FSP1-CoQ₁₀. Therefore, after its reaction with a lipid-derived peroxy radical, α-tocopherol could either be regenerated by reduced CoQ₁₀ or directly in vitro by FSP1 without the need for CoQ₁₀ (Extended Data Fig. 5h–j). Direct monitoring of phospholipid hydroperoxide formation in linoleate-rich liposomes corroborated the results of the co-autoxidation experiments, showing

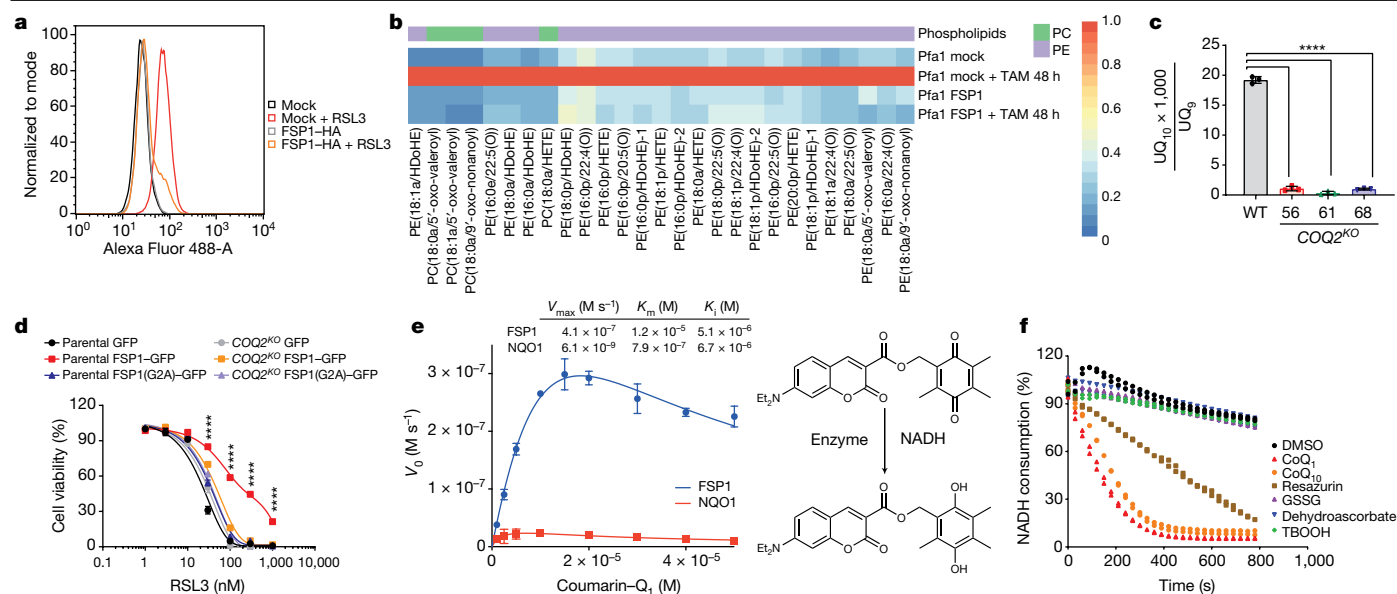


Fig. 3 | FSP1 protects cells against unrestrained lipid peroxidation. **a**, Flow cytometry analysis of RSL3-induced (300 nM for 3 h) BODIPY 581/591 C11 oxidation in Pfa1 cells overexpressing mock or FSP1-HA. Data show one representative of two independently performed experiments. **b**, Heat map showing the representation of mono-oxidized phospholipid species (PE, phosphatidylethanolamines; PC, phosphatidylcholine) in mock and FSP1-HA-expressing Pfa1 cells treated with or without 4-hydroxytamoxifen (TAM) for 48 h. For the heat map, samples ($n = 6$) were averaged and normalized to cell number (1×10^6 cells). Each lipid species was normalized to the maximum detected level. The experiment was performed independently twice. **a**, acyl; e, plasmalyl; p, plasmenyl/plasmalogen. **c**, Relative quantification of ubiquinone CoQ₁₀ ($[M + NH_4]^+$ $m/z = 880.7177$, retention time = 22.8 min) in parental HT1080 and COQ2-knockout HT1080 clones using liquid chromatography–mass spectrometry. Ubiquinone 9 ($[M + NH_4]^+$ $m/z = 812.6551$, retention time = 12.3 min) was used as internal standard. **d**, Dose-dependent toxicity of RSL3 in parental and COQ2-knockout HT1080

cells overexpressing FSP1-GFP, FSP1(G2A)-GFP or GFP. Cell viability was assessed after 24 h using Aquabluer. **e**, Kinetic parameters for the reduction of coumarin–quinone (Q₁) conjugate by FSP1 (50 nM, blue) and NQO1 (50 nM, red) in Tris-buffered saline (10 mM, pH 7.4) in the presence of NADH (200 μ M) at 37 °C. Initial rates were determined from the fluorescence of the product hydroquinone (excitation, 415 nm; emission, 470 nm). The data are fitted to a standard substrate inhibition model and are mean \pm s.d. **f**, NADH consumption assay (340 nm) in TBS buffer using recombinant purified human FSP1 in combination with different electron acceptor molecules (ubiquinone-1 (CoQ₁), ubiquinone-10 (CoQ₁₀), resazurin, oxidized glutathione (GSSG), dehydroascorbate and tert-butyl hydroperoxide (TBOOH)). Data represent $n = 2$ technical replicates of one out of three independent experiments (**f**). Data are mean \pm s.d. of $n = 4$ (**d**) or $n = 3$ (**c**, **e**) wells of a 96-well plate from one representative of three (**e**) or one (**c**, **d**) independent experiments, **** $P < 0.0001$; one-way ANOVA (**c**) and two-way ANOVA (**d**).

substantial FSP1-catalysed suppression of pLPO that was further enhanced in the presence of both CoQ₁₀ and α -tocopherol (Extended Data Fig. 5k).

Loss of FSP1 sensitizes to ferroptosis

On the basis of the strong protective effect provided by FSP1 and the possibility to maintain cells in the absence of GPX4, we imagined that a counter-screen of FSP1-overexpressing cells in a GPX4 knockout or wild-type background could be useful for the discovery of FSP1 inhibitors. We screened approximately 10,000 drug-like compounds⁴, which led to the identification of iFSP1 as a potent FSP1 inhibitor (Fig. 4a). iFSP1 selectively induced ferroptosis in GPX4-knockout Pfa1 and HT1080 cells that overexpressed FSP1 (Extended Data Fig. 6a, b). Preliminary structure–activity relationship studies have yet to identify compounds with substantial improvement over iFSP1 (Extended Data Fig. 6c).

To determine whether FSP1 could serve as a ferroptosis suppressor in cancer, we generated a monoclonal antibody against human FSP1 (Extended Data Fig. 6d), and explored its expression along with the main ferroptosis players in a panel of human cancer cell lines of different origins (Extended Data Fig. 7). Indeed, FSP1 was expressed in most tumour cell lines, and iFSP1 treatment robustly sensitized these cells to RSL3-induced ferroptosis (Extended Data Fig. 8). We then generated FSP1-knockout and FSP1-overexpressing cells from a selection of these cell lines (Fig. 4b, c and Extended Data Fig. 7) and compared the effects of pharmacological inhibition (iFSP1) and FSP1

knockout on the sensitization of cells to ferroptosis. As expected, genetic deletion of FSP1 was more efficient than small-molecule inhibition, whereas iFSP1 treatment in the FSP1-knockout background had no additive effect to RSL3-induced ferroptosis (Extended Data Fig. 6e, f). Notably, a few cells that were sensitive to RSL3 could not be resensitized by iFSP1 when FSP1 was overexpressed. This may be due to drug metabolism and excretion, and these effects should be investigated further (Extended Data Fig. 6f). Detailed experiments demonstrated that FSP1 knockout in MDA-MB-436 cells lowered their resistance to RSL3-induced ferroptosis, whereas mouse FSP1 re-expression restored the resistance of cells to ferroptosis (Fig. 4d, e). Analysis of the cancer dependency map (DepMap; <https://depmap.org/portal/>) revealed that lower expression of FSP1 correlates with an increased GPX4 dependency in a panel of 559 cancer cell lines (Extended Data Fig. 9a). Additionally, FSP1 expression directly correlated with resistance to ferroptosis inducers RSL3, ML162 and ML210 in a panel of 860 cancer cell lines (<https://portals.broadinstitute.org/ctdp>) (Extended Data Fig. 9b). No synergistic cell death was detected with cisplatin or other known cytotoxic compounds (Extended Data Fig. 9c, d), suggesting that FSP1 inhibition selectively sensitizes cells to ferroptosis inducers. This finding is particularly important as therapy-resistant tumours only respond to complete elimination of GPX4 activity; minute amounts are sufficient to sustain cell viability²⁷. Moreover, pharmacological targeting of GPX4 may only achieve partial anti-tumour effects. In fact, in mice bearing human xenografts, a companion study¹⁸ demonstrates that the growth of H460 tumours can only be reduced by concomitant deletion

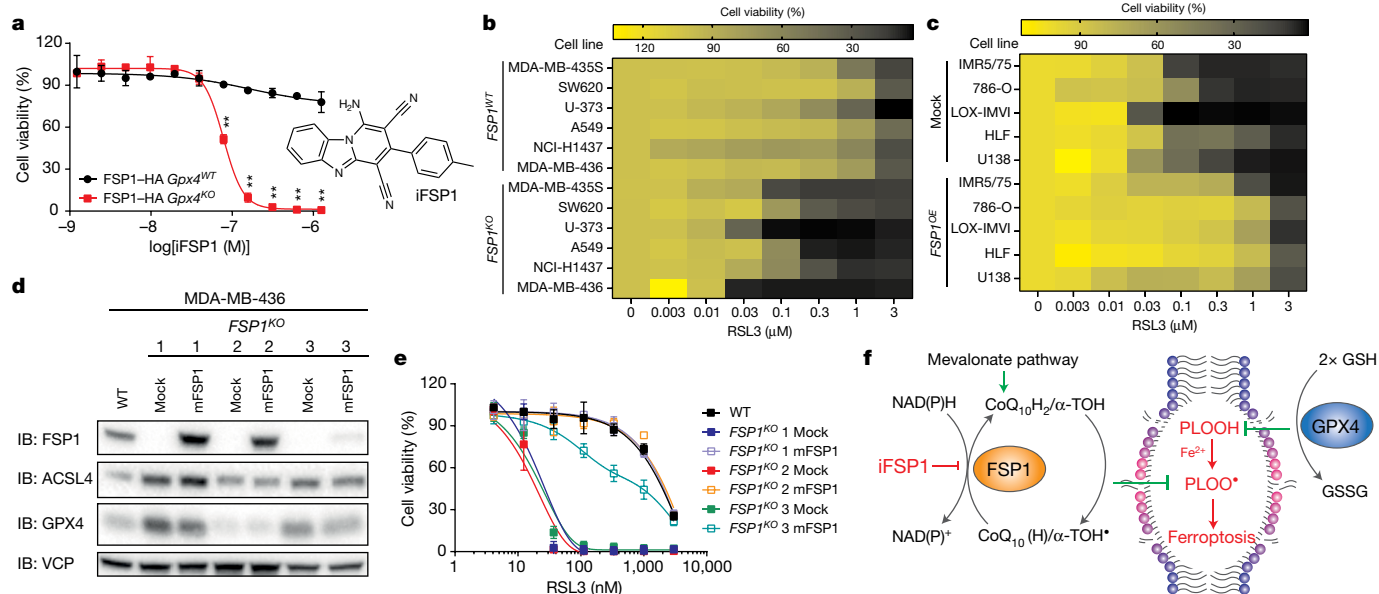


Fig. 4 | FSP1 inhibition sensitizes tumour cells to ferroptosis. a, Chemical structure of iFSP1. Dose-dependent toxicity of iFSP1 in wild-type and *Gpx4*-knockout Pfa1 cells overexpressing FSP1-HA. **b, c**, Heat maps depicting the dose-dependent toxicity of RSL3 in a panel of genetically engineered human cancer cell lines (*FSP1* knockout (**b**); *FSP1* overexpression (OE) (**c**); for detailed cell viability assays including iFSP1 and lipoxstatin-1 treatments, see Extended Data Fig. 6e, f). **d**, Immunoblot analysis of FSP1, ACSL4, GPX4 and VCP expression in parental MDA-MB-436 cells and three independent *FSP1*-knockout clones (KO1–3) overexpressing mock or mouse FSP1 (mFSP1). Immunoblot images are cropped from the chemiluminescence signal files. For

gel source data showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1. **e**, Dose-dependent toxicity of RSL3 of the cell lines depicted in **d**. Expression of FSP1 restored resistance to RSL3-induced ferroptosis in all three clones. **f**, Graphical abstract depicting the anti-ferroptotic function of FSP1 as a glutathione-independent suppressor of phospholipid peroxidation by inhibition of lipid radical-mediated autoxidation, initiated by peroxy radicals (PLOO[•]), of lipid bilayers. Data are mean \pm s.d. of $n = 3$ wells of a 96-well plate from one representative of one (**a**) or two (**b, c, e**) independent experiments; ** $P < 0.01$; two-way ANOVA.

of GPX4 and FSP1, whereas GPX4 single-knockout tumours grow normally.

Our data establish that the NADH–FSP1–CoQ₁₀ pathway is a potent suppressor of pLPO and ferroptosis (Fig. 4f). As such, phospholipid redox homeostasis can be disassociated from the glutathione–GPX4 axis, and can be further exploited pharmacologically to efficiently sensitize cancer cells to ferroptosis inducers. Our discovery explains why NAD(P)H²⁸ and defects in the mevalonate pathway through loss of ubiquinone^{13,29} converge on FSP1 and thereby predict sensitivity to ferroptosis. Furthermore, our data provide a compelling case for the long-debated antioxidant role^{30,31} of extra-mitochondrial CoQ₁₀ and suggest that its beneficial effects should be investigated further alongside FSP1.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1707-0>.

- Dixon, S. J. et al. Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* **149**, 1060–1072 (2012).
- Conrad, M., Angeli, J. P., Vandenabeele, P. & Stockwell, B. R. Regulated necrosis: disease relevance and therapeutic opportunities. *Nat. Rev. Drug Discov.* **15**, 348–366 (2016).
- Yang, W. S. et al. Regulation of ferroptotic cancer cell death by GPX4. *Cell* **156**, 317–331 (2014).
- Friedmann Angeli, J. P. et al. Inactivation of the ferroptosis regulator Gpx4 triggers acute renal failure in mice. *Nat. Cell Biol.* **16**, 1180–1191 (2014).
- Zilka, O. et al. On the mechanism of cytoprotection by ferrostatin-1 and lipoxstatin-1 and the role of lipid peroxidation in ferroptotic cell death. *ACS Cent. Sci.* **3**, 232–243 (2017).

- Shah, R., Shchepinov, M. S. & Pratt, D. A. Resolving the role of lipoygenases in the initiation and execution of ferroptosis. *ACS Cent. Sci.* **4**, 387–396 (2018).
- Stockwell, B. R. et al. Ferroptosis: a regulated cell death nexus linking metabolism, redox biology, and disease. *Cell* **171**, 273–285 (2017).
- Tarangelo, A. et al. p53 suppresses metabolic stress-induced ferroptosis in cancer cells. *Cell Rep.* **22**, 569–575 (2018).
- Kagan, V. E. et al. Oxidized arachidonic and adrenic PEs navigate cells to ferroptosis. *Nat. Chem. Biol.* **13**, 81–90 (2017).
- Doll, S. et al. ACSL4 dictates ferroptosis sensitivity by shaping cellular lipid composition. *Nat. Chem. Biol.* **13**, 91–98 (2017).
- Wu, M., Xu, L. G., Li, X., Zhai, Z. & Shu, H. B. AMID, an apoptosis-inducing factor-homologous mitochondrion-associated protein, induces caspase-independent apoptosis. *J. Biol. Chem.* **277**, 25617–25623 (2002).
- Ingold, I. et al. Selenium utilization by GPX4 is required to prevent hydroperoxide-induced ferroptosis. *Cell* **172**, 409–422 (2018).
- Viswanathan, V. S. et al. Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* **547**, 453–457 (2017).
- Hangauer, M. J. et al. Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature* **551**, 247–250 (2017).
- Tsoi, J. et al. Multi-stage differentiation defines melanoma subtypes with differential vulnerability to drug-induced iron-dependent oxidative stress. *Cancer Cell* **33**, 890–904 (2018).
- Angeli, J. P. F., Shah, R., Pratt, D. A. & Conrad, M. Ferroptosis inhibition: mechanisms and opportunities. *Trends Pharmacol. Sci.* **38**, 489–498 (2017).
- Horikoshi, N., Cong, J., Kley, N. & Shenk, T. Isolation of differentially expressed cDNAs from p53-dependent apoptotic cells: activation of the human homologue of the *Drosophila* peroxidase gene. *Biochem. Biophys. Res. Commun.* **261**, 864–869 (1999).
- Bersuker, K. et al. The CoQ oxidoreductase FSP1 acts parallel to GPX4 to inhibit ferroptosis. *Nature* <https://doi.org/10.1038/s41586-019-1705-2> (2019).
- Seiler, A. et al. Glutathione peroxidase 4 senses and translates oxidative stress into 12/15-lipoxygenase dependent- and AIF-mediated cell death. *Cell Metab.* **8**, 237–248 (2008).
- Eisenhaber, F. et al. Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-PI, NMT and PTS1. *Nucleic Acids Res.* **31**, 3631–3634 (2003).
- Borgese, N., Aggujaro, D., Carrera, P., Pietrini, G. & Bassetti, M. A role for N-myristoylation in protein targeting: NADH-cytochrome b₅ reductase requires myristic acid for association with outer mitochondrial but not ER membranes. *J. Cell Biol.* **135**, 1501–1513 (1996).
- Mousnier, A. et al. Fragment-derived inhibitors of human N-myristoyltransferase block capsid assembly and replication of the common cold virus. *Nat. Chem.* **10**, 599–606 (2018).

23. Elguindy, M. M. & Nakamaru-Ogiso, E. Apoptosis-inducing factor (AIF) and its family member protein, AMID, are rotenone-sensitive NADH:ubiquinone oxidoreductases (NDH-2). *J. Biol. Chem.* **290**, 20815–20826 (2015).
24. Frei, B., Kim, M. C. & Ames, B. N. Ubiquinol-10 is an effective lipid-soluble antioxidant at physiological concentrations. *Proc. Natl Acad. Sci. USA* **87**, 4879–4883 (1990).
25. Haidasz, E. A., Van Kessel, A. T. & Pratt, D. A. A continuous visible light spectrophotometric approach to accurately determine the reactivity of radical-trapping antioxidants. *J. Org. Chem.* **81**, 737–744 (2016).
26. Niki, E. Mechanisms and dynamics of antioxidant action of ubiquinol. *Mol. Aspects Med.* **18**, 63–70 (1997).
27. Mannes, A. M., Seiler, A., Bosello, V., Maiorino, M. & Conrad, M. Cysteine mutant of mammalian GPx4 rescues cell death induced by disruption of the wild-type selenoenzyme. *FASEB J.* **25**, 2135–2144 (2011).
28. Shimada, K., Hayano, M., Pagano, N. C. & Stockwell, B. R. Cell-line selectivity improves the predictive power of pharmacogenomic analyses and helps identify NADPH as biomarker for ferroptosis sensitivity. *Cell Chem. Biol.* **23**, 225–235 (2016).
29. Shimada, K. et al. Global survey of cell death mechanisms reveals metabolic regulation of ferroptosis. *Nat. Chem. Biol.* **12**, 497–503 (2016).
30. Morré, D. J. & Morré, D. M. Non-mitochondrial coenzyme Q. *Biofactors* **37**, 355–360 (2011).
31. Nyquist, S. E., Barr, R. & Morré, D. J. Ubiquinone from rat liver Golgi apparatus fractions. *Biochim. Biophys. Acta* **208**, 532–534 (1970).
32. Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
33. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
34. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

For immunoblot source data, see Supplementary Fig. 1. Source Data for Figs. 1–4 and Extended Data Figs. 1–6, 8, 9 are provided with the paper.

Acknowledgements This work is supported by the Junior Group Leader program of the Rudolf Virchow Center, University of Würzburg and Deutsche Forschungsgemeinschaft (DFG) FR 3746/3-1 to J.P.F.A., the DFG CO 291/5-2 and CO 291/7-1, the German Federal Ministry of Education and Research (BMBF) through the Joint Project Modelling ALS Disease In vitro (MAIV, 01EK1611B) and the VIP+ program NEUROPROTEKT (03VP04260), as well as the m⁴ Award provided by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) to M.C., the Cancer Research UK (CRUK, grants C29637/A20183 and C29637/A21451) to E.W.T., the European Research Council (LipidArrays) to V.O., the Natural Sciences and Engineering Council of Canada and the Canada Foundation for Innovation to D.A.P. and PhD scholarship by DFG GRK2157 to A.K.

Author contributions M.C., J.P.F.A. and S.D. conceived the study and wrote the manuscript. M.A. and V.O. performed (oxi)lipidomics analysis and data interpretation. S.D., B.P., E.P., D.W., F.P.F., J.P.F.A., T.V., V.M., I.I., K.B., M. Sato, M.R., T.N.X.d.S. and M.C.d.S. performed in vitro experiments. R.S. and D.A.P. performed functional characterization of recombinant FSP1. S.D., F.P.F., D.A.P., J.P.F.A. and M.C. performed evaluation and interpretation of the in vitro data. M. Sattler, A.M. and G.M.P. expressed and purified recombinant FSP1. C.H.S. provided TNBC cell lines. A.F. and A. Schepers helped to generate the monoclonal antibodies. B.P. and J.W. carried out screening of FSP1 inhibitors and related structure–activity relationship studies. W.S. and A. Schulze performed liquid chromatography–mass spectrometry analysis of ubiquinone content. A.G.G. and E.W.T. studied myristoylation of FSP1. A.K., M. Sauer, F.P.F. and J.P.F.A. performed enhanced microscopy experiments. All authors read and agreed on the content of the paper.

Competing interests The authors declare no competing interests.

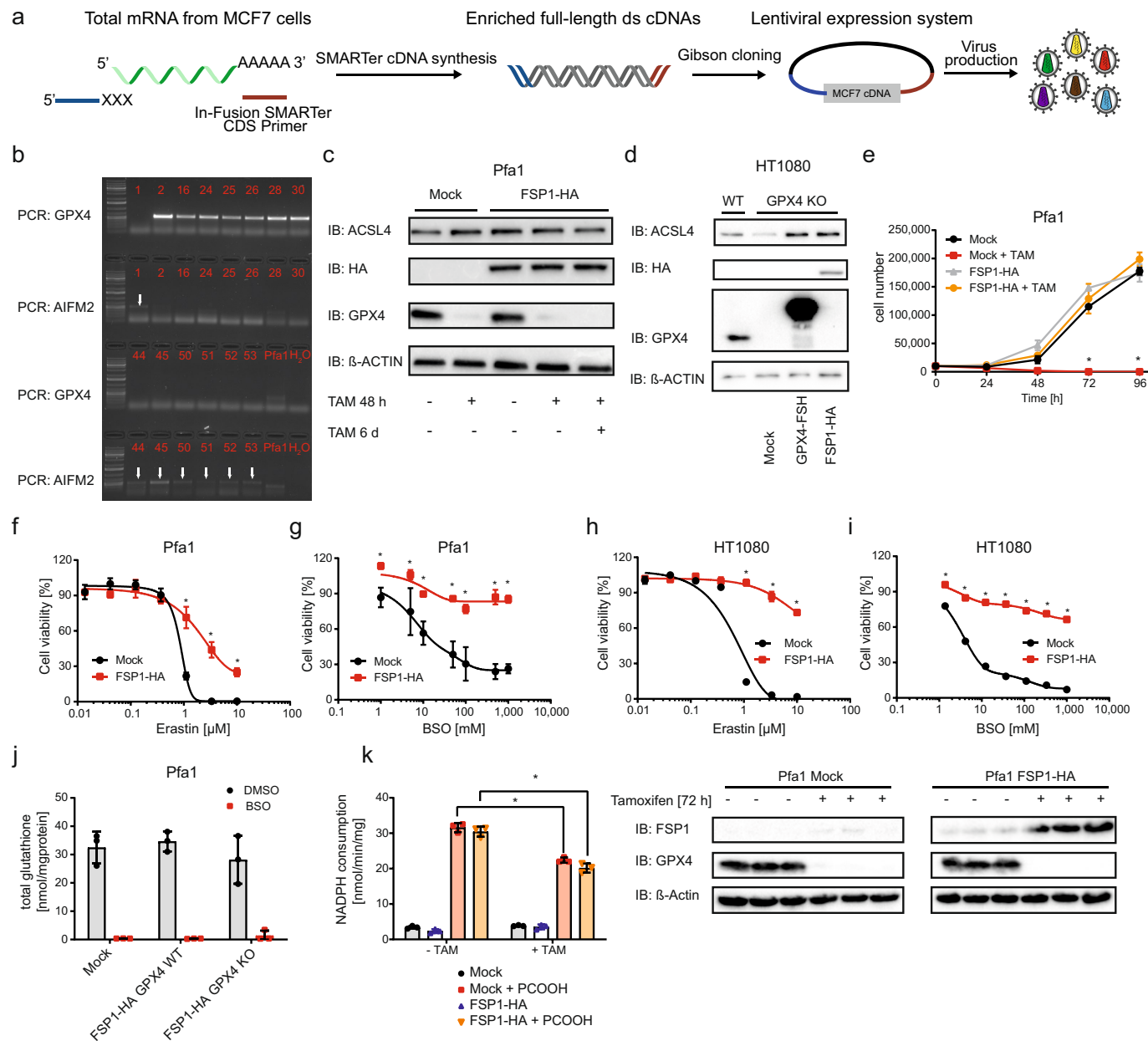
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1707-0>.

Correspondence and requests for materials should be addressed to J.P.F.A. or M.C.

Peer review information *Nature* thanks Kivanc Birsoy, Navdeep S. Chandel and Brent R. Stockwell for their contribution to the peer review of this work.

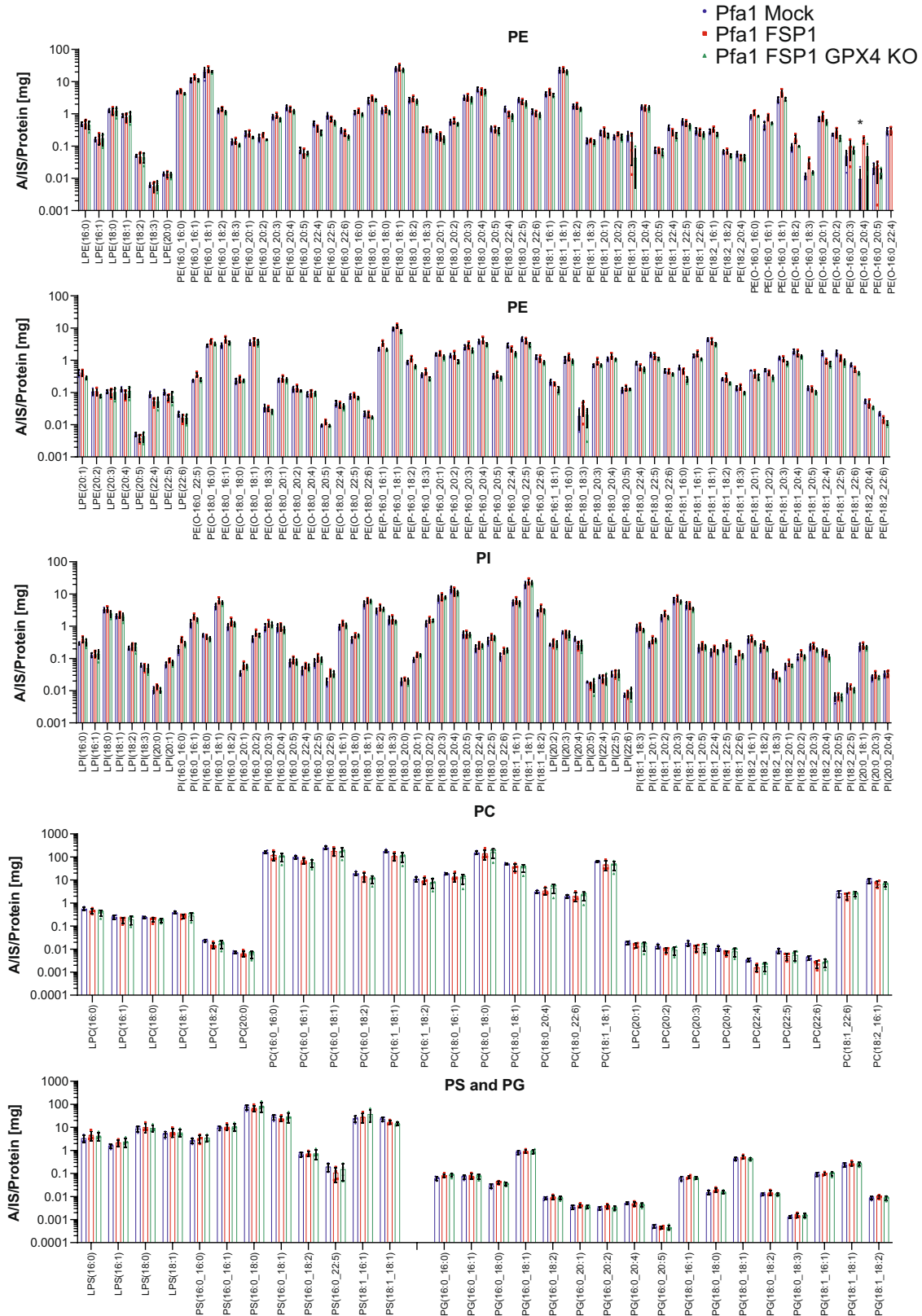
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Identification and characterization of FSP1 as an anti-ferroptotic protein. a, Schematic depicting the generation of a lentiviral cDNA-overexpression library using the total mRNA from MCF7 cells.

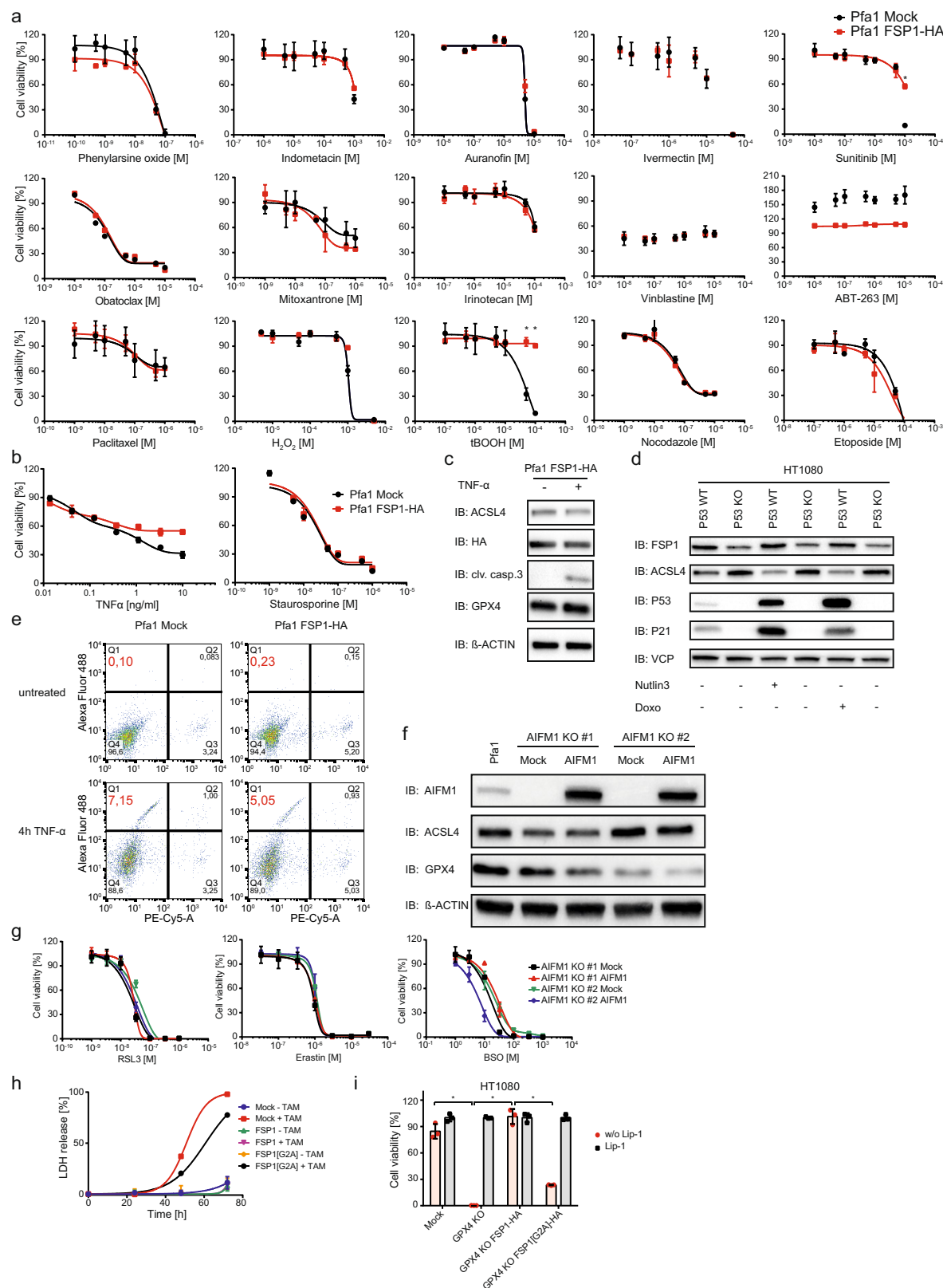
b, Genomic PCRs of the 14 Pfa1 cell clones that remained clones after the removal of false-positive results using human-specific primers to amplify the human cDNAs of *GPX4* (571 bp) or *AIFM2* (524 bp). The clones 2, 16, 24, 25, 26, 28 and 30 showed positive PCR results for *GPX4* (571 bp). Clones 1, 44, 45, 50, 51, 52 and 53 were positive for *AIFM2* (524 bp) as indicated by the white arrows. Data are one of $n=3$ independent experiments. **c**, Immunoblot analysis of ACSL4, HA, GPX4 and β -actin expression in Pfa1 cells stably expressing mock or FSP1-HA. *Gpx4* deletion was induced by the administration of TAM for the indicated time period. **d**, Immunoblot analysis of ACSL4, HA, GPX4 and β -actin expression in wild-type and *GPX4*-knockout HT1080 cells stably expressing mock, GPX4-FSH or FSP1-HA. **e**, Proliferation of mock and FSP1-HA Pfa1 cells treated with or without TAM. Cell numbers were assessed every 24 h using a Neubauer haemocytometer. Data are mean \pm s.d. of $n=3$ wells of a 24-well plate from one representative of two independent experiments. **f, g**, Dose-dependent toxicity of erastin (**f**) and L-buthionine sulfoximine (BSO; **g**), which is an inhibitor of γ -glutamyl-cysteine ligase, in Pfa1 cells expressing mock or

FSP1-HA. **h, i**, Dose-dependent toxicity of erastin (**h**) and BSO (**i**) in HT1080 cells expressing mock or FSP1-HA. Cell viability was assessed 48 h (**f, h**) or 72 h (**g, i**) after treatments using Aquabluer. Data are mean \pm s.d. of $n=3$ wells of a 96-well plate from one representative of three (**f-i**) independent experiments. * $P<0.01$; two-way ANOVA. **j**, Measurement of total glutathione levels in Pfa1 mock, FSP1-expressing and FSP1-expressing *Gpx4*-knockout cells treated with or without BSO. Data are mean \pm s.d. of $n=3$ wells of a 96-well plate from one representative of three independent experiments. **k**, Left, determination of NADPH consumption by glutathione reductase as an indirect measure of the GPX4 activity. Phosphatidylcholine lipid hydroperoxide (PCOOH) was used as a GPX4-specific substrate. Right, cell lysates from mock and FSP1-HA Pfa1 cells treated with or without TAM for 48 h were used for the assay as shown by the immunoblot (FSP1, GPX4 and β -actin). FSP1 was detected using the polyclonal antibody (PAS-24562). Data are mean \pm s.d. of $n=3$ wells of a 6-well plate from one representative of three independent experiments. Immunoblot images (**c, d, k**) are cropped from the chemiluminescence signal files. For gel source data (**c, d, k**) showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1.



Extended Data Fig. 2 | FSP1 expression does not change the phospholipid composition. Lipidomic profile (only detectable phospholipid species) of phosphatidylethanolamine (PE), phosphatidylcholine (PC), phosphatidylglycerol (PG), phosphatidylinositol (PI) and phosphatidylserine (PS), including plasmalogen (O) and plasmalogen (P) lipids of mock, FSP1-HA and *Gpx4*-knockout FSP1-HA Pfa1 cells. Data are the mean values of the area of

analyte (A) over the internal standard (IS) per total protein (mg) of $n = 4$ replicates of one experiment performed independently three times. \log_{10} -transformation has been applied to better visualize and compare the abundance of the different phospholipid species in the samples. * $P < 0.05$; multiple t -test with Sidak–Bonferroni correction for multiple comparisons.

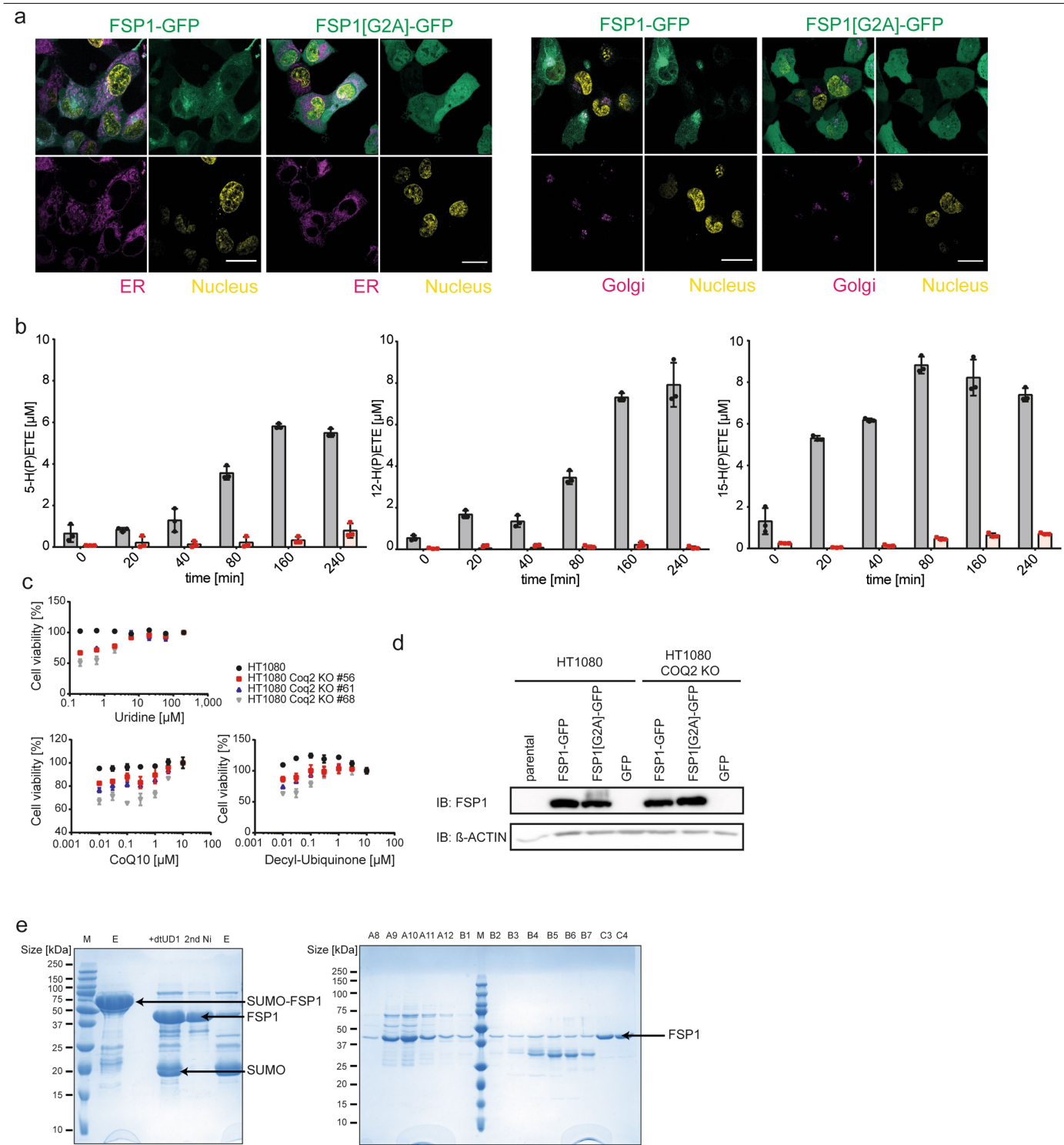


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | FSP1 is a highly specific anti-ferroptotic protein.

a, Dose-dependent toxicity of phenylarsine oxide, indomethacin, auranofin, ivermectin, sunitinib, obatoclox, mitoxantrone, irinotecan, vinblastine, ABT-263, nocodazole, etoposide, paclitaxel, H₂O₂ and tert-butyl hydroperoxide (tBOOH) of Pfa1 cells expressing mock or FSP1-HA. Cell viability was assessed 24 h after treatment using Aquabluer. **b**, Dose-dependent toxicity of TNF and staurosporine of mock and FSP1-HA-expressing Pfa1 cells. Cell viability was assessed 24 h after treatment using Aquabluer. **c**, Immunoblot analysis (ACSL4, HA, cleaved caspase 3 (clv. Casp3), GPX4 and β -actin) of Pfa1 FSP1-HA cells treated with or without TNF for 6 h. **d**, Immunoblot analysis of FSP1, ACSL4, p53, p21 and VCP expression in *p53* (also known as *TP53*) wild-type and *p53*-knockout (CRISPR-CAS9-modified) HT1080 cell lines treated with the MDM2 (*MDM2* proto-oncogene) inhibitor Nutlin3 or the cytostatic compound doxorubicin (Doxo). Expression of FSP1 was not altered by Nutlin3 or doxorubicin treatment, whereas the expression of p53 and p21 was strongly induced in HT1080 *p53* wild-type cells. Data show one representative of *n* = 3 independent experiments. **e**, Flow cytometry analysis of annexin V/propidium iodide staining in Pfa1 cells expressing mock or FSP1-HA treated with or without TNF for 4 h. No difference in the apoptotic activity was observed in cells as visualized in the Alexa Fluor 488/PE-Cy5 channels. Data show one

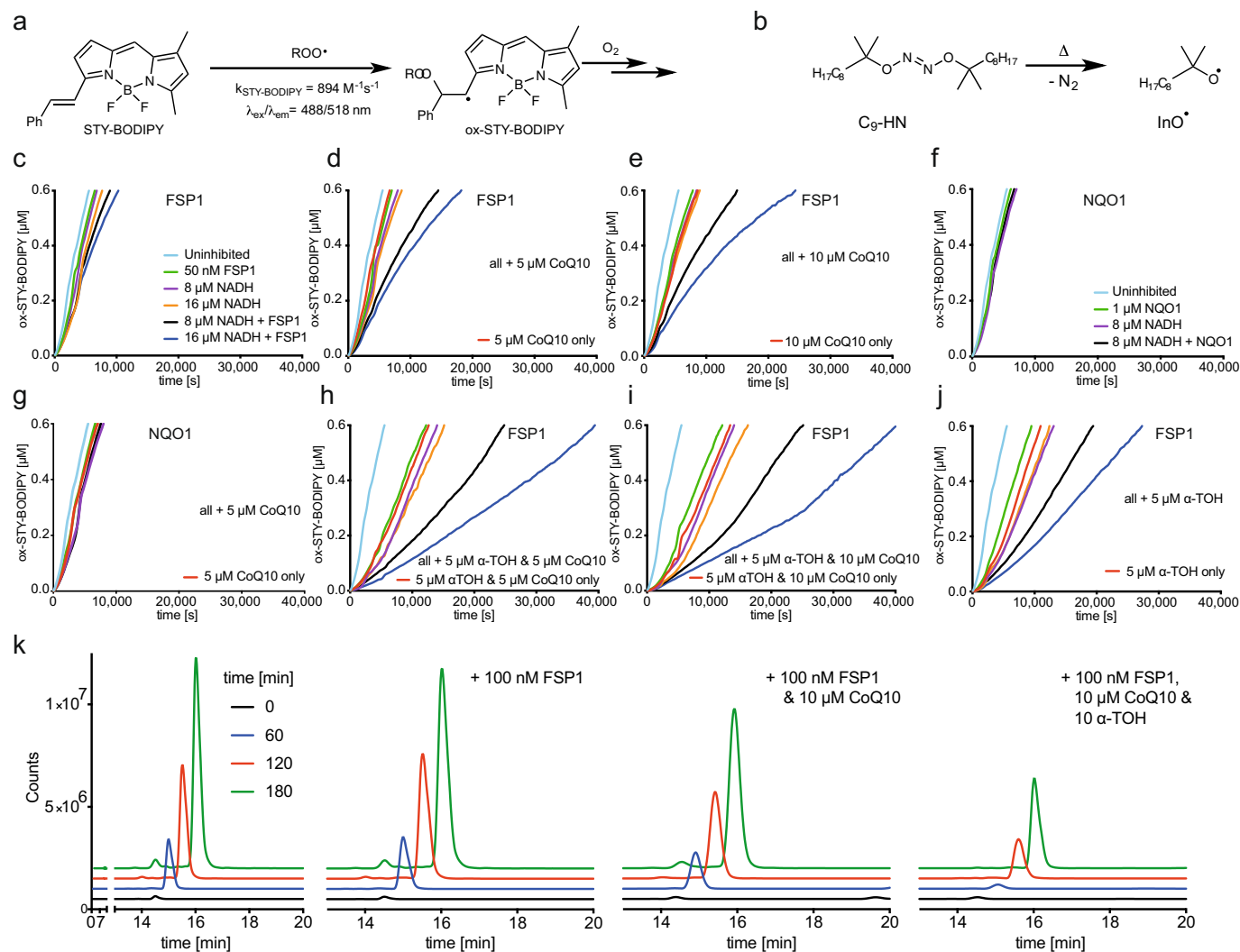
representative experiment of an experiment performed independently twice. **f**, Immunoblot analysis of AIFM1, ACSL4, GPX4 and β -actin in two different Pfa1 *Aifm1*-knockout cell clones overexpressing mock or AIFM1. Data show one representative of *n* = 3 independent experiments. **g**, Dose-dependent toxicity of RSL3, erastin and BSO in *Aifm1*-knockout Pfa1 cell clones (1 and 2) overexpressing mock or AIFM1. AIFM1 expression does not affect ferroptosis sensitivity. Data are the mean of *n* = 3 replicates of a representative experiment performed independently three times. **h**, Time-dependent lactate dehydrogenase (LDH) release of Pfa1 cells stably expressing mock, FSP1-HA or FSP1(G2A) treated with TAM to induce GPX4 loss. Supernatants were collected from 6-well plates at different time points after TAM induction and assayed for lactate dehydrogenase content in a 96-well plate. **i**, Wild-type and *GPX4*-knockout HT1080 cells overexpressing mock, hGPX4-FSH, FSP1-HA or FSP1(G2A)-HA treated with and without 200 nM Lip-1. Cell viability was assessed after 72 h using Aquabluer. Data are the mean \pm s.d. of *n* = 3 wells of a 96-well plate from one representative of three independent experiments (**a**, **b**, **g**–**i**); **P* < 0.01; two-way ANOVA. Immunoblot images (**c**, **d**, **f**) are cropped from the chemiluminescence signal files. For gel source data (**c**, **d**, **f**) showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1.



Extended Data Fig. 4 | See next page for caption.

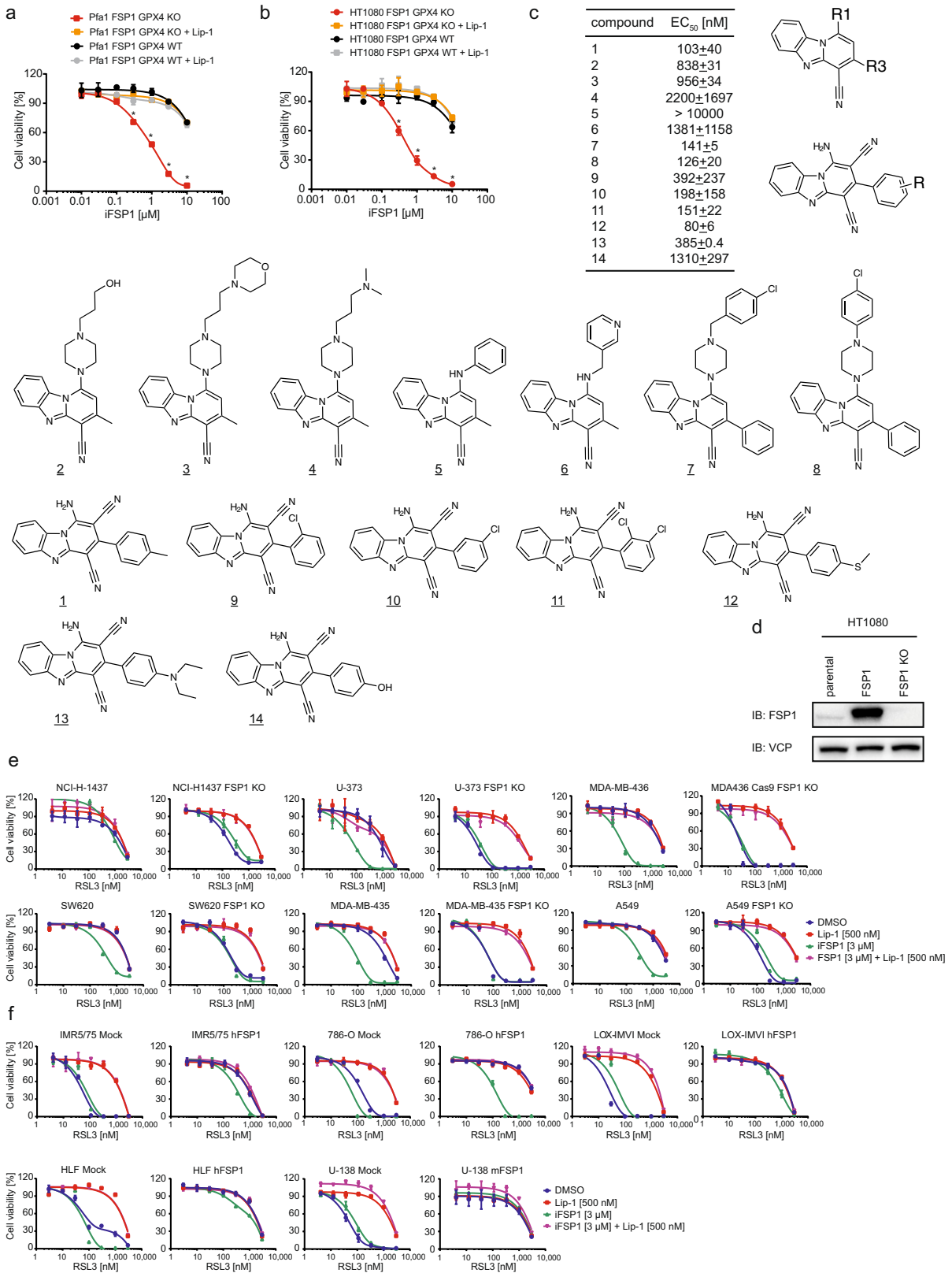
Extended Data Fig. 4 | FSP1 protects against unrestrained lipid peroxidation in a COQ2-dependent manner. **a**, Enhanced resolution confocal microscopy images demonstrating different localizations of FSP1-GFP and the FSP1(G2A)-GFP mutant in HT1080 cells. DAPI (yellow), GFP (green), endoplasmic reticulum or Golgi tracker (magenta). Scale bars, 20 nm. Data show one representative of $n = 3$ independently performed experiments. **b**, Formation of 5-hydro(pero)xyeicosatetraenoic acid (5-H(P)ETE) (multiple reaction monitoring (MRM): 319 \rightarrow 115), 12-H(P)ETE (MRM: 319 \rightarrow 179) and 15-H(P)ETE (MRM: 319 \rightarrow 219) in either mock (black) or FSP1-HA-overexpressing (red) Pf1 cells treated with 0.2 μ M RSL3 and 40 μ M Arachidonic acid. Hydroperoxides were analysed as their alcohols following reduction with PPH₃ (triphenylphosphane) in methanol. Data are the mean of biological triplicates from one representative of $n = 3$ independently performed experiments. **c**, Dose-dependent rescue of three independent COQ2-knockout HT1080 cell clones (56, 61 and 68) by supplementation of the cell culture medium with uridine, CoQ₁₀ or decyl-ubiquinone. Cell viability was assayed using the Aquabluer assay 48 h after treatment. Data are mean \pm s.d. of $n = 3$ wells of a 96-well plate performed once.

d, Immunoblot analysis of FSP1 and β -actin in HT1080 parental (left) and HT1080 COQ2-knockout (56) (right) cells overexpressing FSP1-GFP, FSP1(G2A)-GFP or GFP. Immunoblot images are cropped from the chemiluminescence signal files. For gel source data showing the uncropped chemiluminescence signals, see Supplementary Fig. 1. **e**, SDS gels showing the different purification steps of recombinant FSP1 from bacterial cell lysates. Left, SDS gel of protein extracts after initial nickel affinity chromatography (E1), the SUMO-tag was cleaved in the eluate by addition of the SUMO protease (dtUD1) and a second round of nickel affinity chromatography was performed to remove the cleaved SUMO-tag as well as uncleaved SUMO-FSP1 and SUMO protease (E2). The flow-through fraction was collected (second nickel). The SUMO-FSP1 fusion protein is visible around 55 kDa and FSP1 at 40.5 kDa. Right, SDS gel showing different fractions containing FSP1 40.5 kDa (A8-A12, B1-B7 and C3-C4) from size-exclusion chromatography of FSP1 after the second nickel-affinity chromatography. Fractions C3 and C4 were used for subsequent assays. One representative of at least three independent experiments.



Extended Data Fig. 5 | FSP1 protects against lipid peroxidation by reducing radical-trapping antioxidants. **a, b**, Co-oxidations of STY-BODIPY (1 μM) (a) and the polyunsaturated lipids of chicken egg phosphatidylcholine liposomes (1 mM). The increase in fluorescence of oxidized STY-BODIPY is monitored over the course of the autoxidation, which is initiated using C₉-HN (0.2 mM) (b). **c**, Representative autoxidations inhibited by 50 nM FSP1 (green), 8 μM NADH (purple), 16 μM NADH (orange), 50 nM FSP1 and 8 μM NADH (black) or 50 nM FSP1 and 16 μM NADH (blue) (c). **d–j**, Analogous representative of inhibited autoxidations to which CoQ₁₀ (d, e), α-tocopherol (α-tocopherol) and

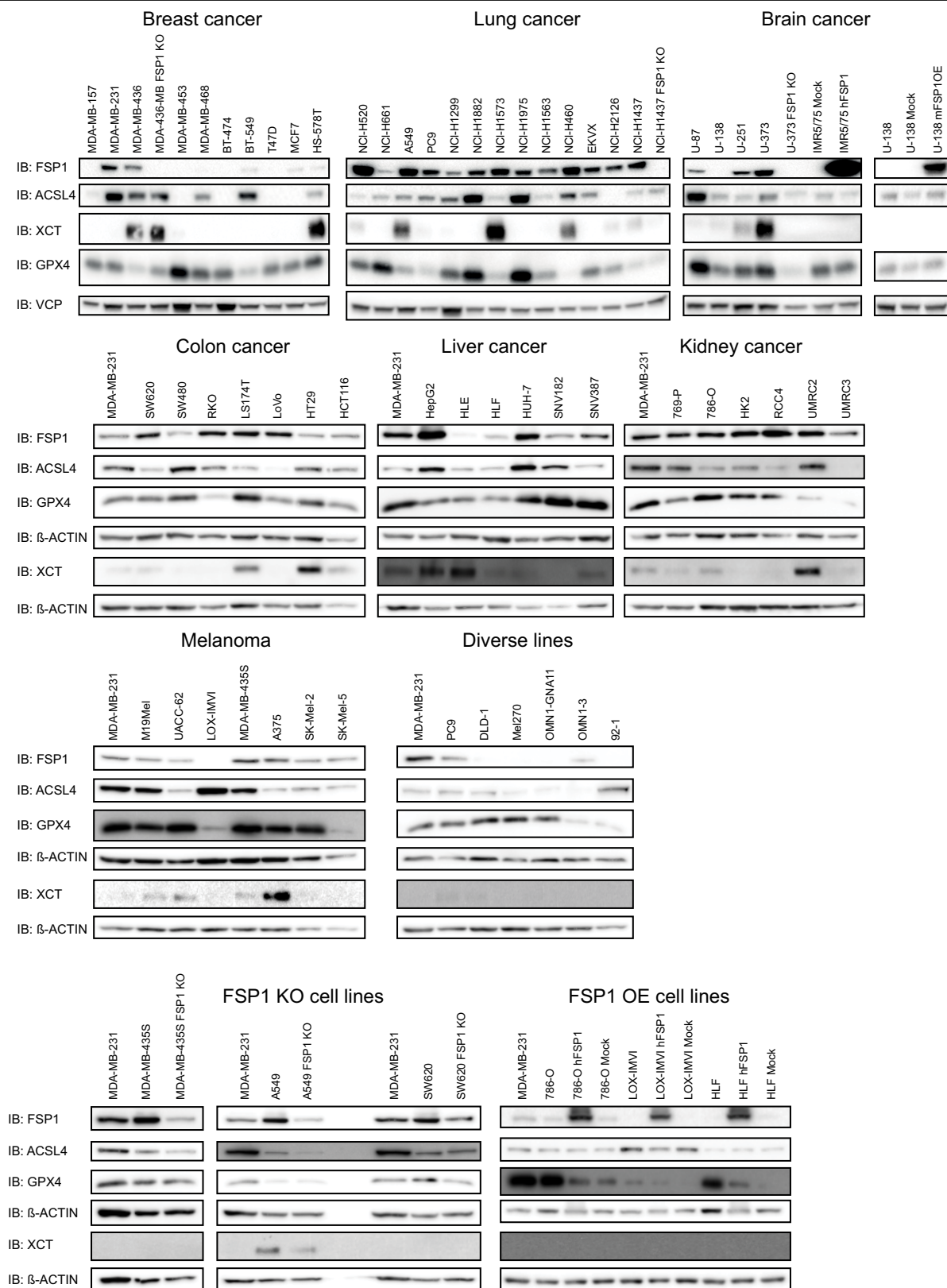
CoQ₁₀ (h, i), or α-tocopherol (j) was added. **f, g**, Recombinant NQO1 failed to suppress autoxidations in a similar manner compared to FSP1 (f, g). **k**, 1-Palmitoyl-2-linoleoyl-phosphatidylcholine hydroperoxide (PLPC-OOH) produced from the autoxidation of soy lecithin liposomes (13.3 mM), inhibited by FSP1 alone, or in the presence of either 10 μM CoQ₁₀ or 10 μM α-tocopherol and 10 μM CoQ₁₀. PLPC-OOH was measured 0, 60, 120 and 180 min after autoxidation was induced using liquid chromatography–mass spectrometry (MRM: 790 → 184). Data show one of $n = 3$ representative experiments.



Extended Data Fig. 6 | See next page for caption.

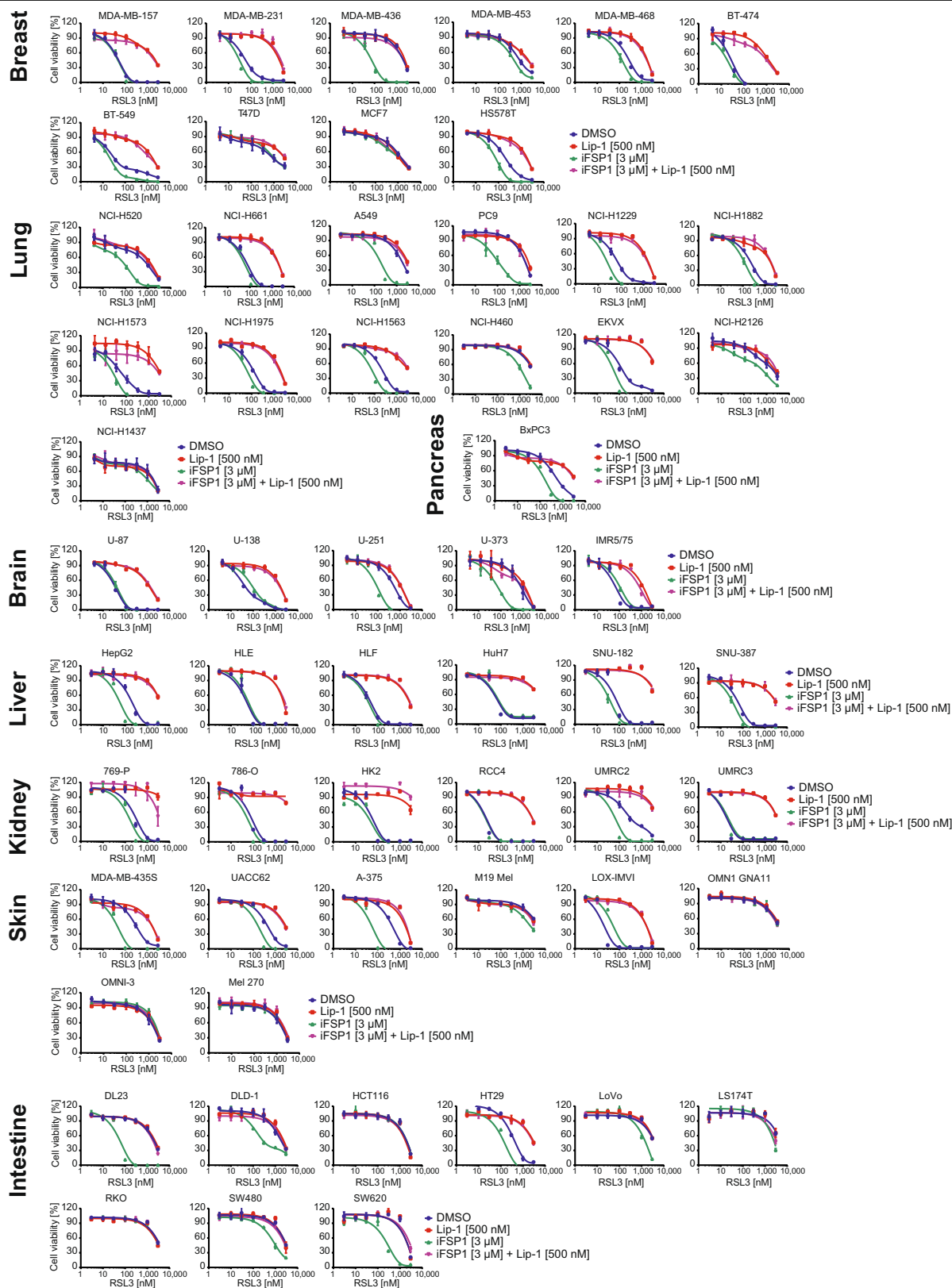
Extended Data Fig. 6 | Development of FSP1-specific inhibitors as ferroptosis sensitizer. a, b, Dose-dependent toxicity of iFSP1 in FSP1-overexpressing cells (Pfa1 (a); HT1080 (b)) with or without *GPX4* loss. Treatment with the ferroptosis inhibitor Lip-1 (150 nM) protected *GPX4*-knockout cells from iFSP1-induced ferroptosis. iFSP1 is only toxic to cells that depend solely (no GPX4 expression detectable) on FSP1 function. **c,** Efficacy of iFSP1 and structurally related analogues; half-maximal effective concentration (EC_{50}) values (mean \pm s.d.) of iFSP1 (**1**) and its derivatives (**2–14**) calculated from experiments performed at least twice in triplicate are shown in the table with the corresponding chemical structures depicted below. Based on commercially available analogues a preliminary structure-activity relationship study revealed that substitution of the amino position (R1, R2) showed broad tolerability of aliphatic groups and that lipophilic substituents of the phenyl group at the 3 position (R3) in the *ortho* and *meta* positions were well tolerated. **d,** Immunoblot analysis of FSP1 and VCP in parental as well as HT1080 FSP1-overexpressing and *FSP1*-knockout HT1080 cells. An in-house-

generated monoclonal antibody against human FSP1 was used. Immunoblot images are cropped from the chemiluminescence signal files. For gel source data showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1. **e,** Dose-dependent toxicity of RSL3 in a panel of genetically engineered (*FSP1*-knockout) human cancer cell lines (NCI-H1437, NCI-H1437 *FSP1* KO, U-373, U-373 *FSP1* KO, MDA-MB-436, MDA-MB-436 *FSP1* KO, SW620, SW620 *FSP1* KO, MDA-MB-435S, MDA-MB-435S *FSP1* KO, A549 and A549 *FSP1* KO) treated with or without FSP1 inhibitor (iFSP1) and Lip-1. **f,** Dose-dependent toxicity of RSL3 in a panel of genetically modified (mouse (mFSP1) and human (hFSP1) *FSP1* overexpression) human cancer cell lines (IMR5/75 mock, IMR5/75 hFSP1, 786-O mock, 786-O hFSP1, LOX-IMVI mock, LOX-IMVI hFSP1, HLF mock, HLF hFSP1, U-138 mock and U-138 mFSP1) treated with or without iFSP1 and Lip-1. Data show the mean \pm s.d. of $n = 3$ wells of a 96-well plate from one representative of three (a–c) or two (e, f) independent experiments; $*P < 0.0001$; two-way ANOVA.



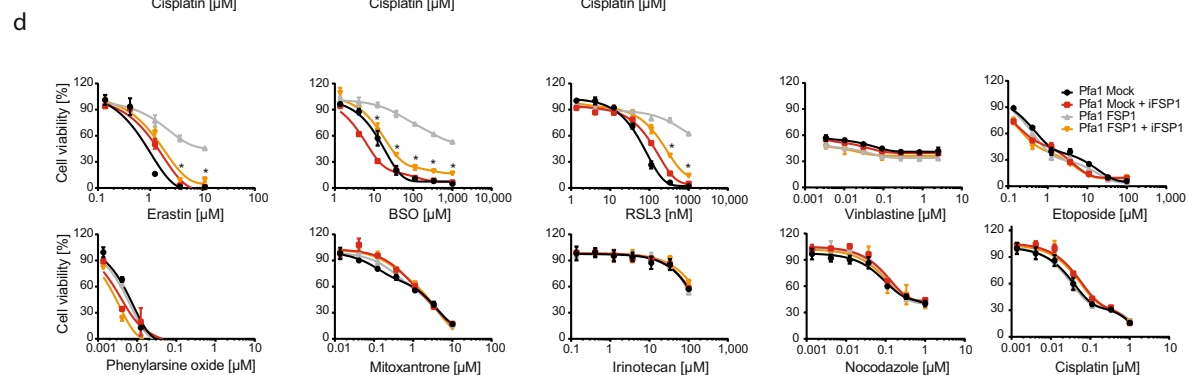
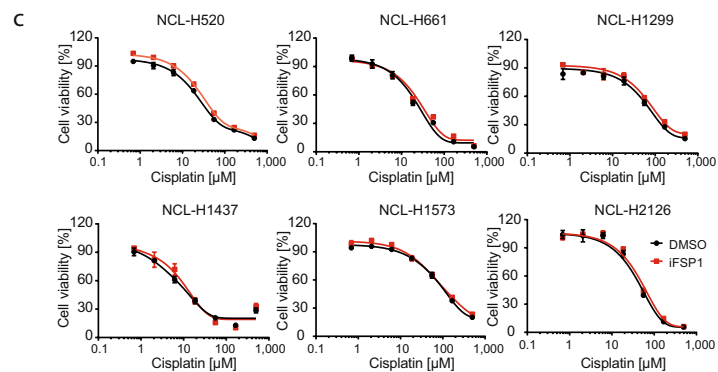
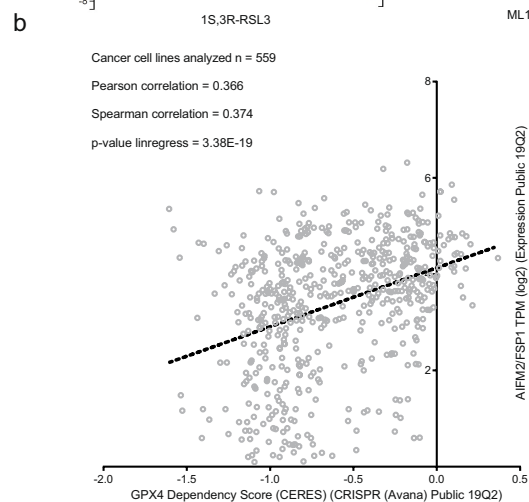
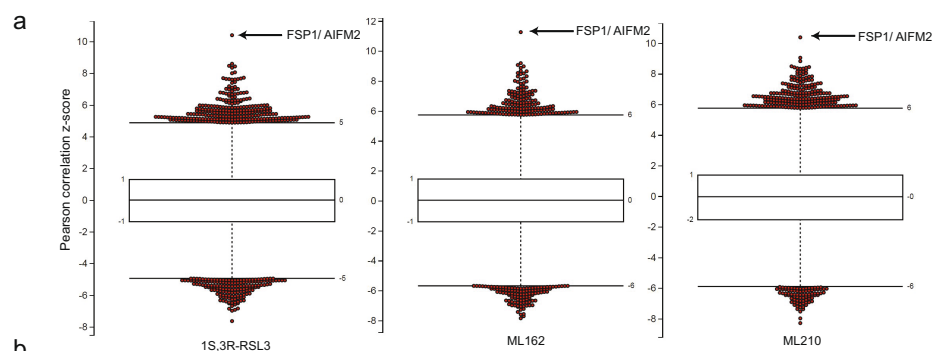
Extended Data Fig. 7 | FSP1 is expressed in a wide range of cancer cell lines.
a, Immunoblot analysis of the expression of key ferroptosis players including ACSL4, FSP1, GPX4 and XCT (SLC7A11) in a panel of cancer cell lines from different origins. In addition, genetically modified cancer cell lines in which FSP1 is knocked out (MDA-436-MB *FSP1* KO, NCI-H1437 *FSP1* KO, U-373 *FSP1* KO, MDA-MB-435S *FSP1* KO, A549 *FSP1* KO and SW620 *FSP1* KO) as well as cell lines with lentiviral overexpression of FSP1 (IMR5/75 hFSP1, 786-O hFSP1, LOX-IMVI

hFSP1 and HLF hFSP1) are shown. VCP or β-actin served as loading control. MDA-MB-231 was used as reference to compare expression levels in between independent blots. Data show one representative of two independent experiments. Immunoblot images are cropped from the chemiluminescence signal files. For gel source data showing the overlap of colorimetric and chemiluminescence signals, see Supplementary Fig. 1.



Extended Data Fig. 8 | iFSP1 sensitizes cancer cell lines from different origins to RSL3-induced ferroptosis. Dose-dependent toxicity of RSL3 in a panel of human cancer cell lines from different origins (breast, lung, pancreas, brain,

liver, kidney, skin and intestine) treated with or without iFSP1 and Lip-1. Data are the mean \pm s.d. of $n = 3$ wells of a 96-well plate from one representative of two independent experiments.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | FSP1 expression directly correlates with resistance to ferroptosis and its inhibition selectively sensitizes cells to ferroptosis.

a, Correlation of a panel of 860 cancer cell lines^{32–34}. The sensitivity to RSL3, ML162 and ML210 was correlated with gene expression. Genes were plotted according to their Pearson correlation score. *FSP1* was the highest ranking gene that correlated with resistance to RSL3 ($P = 0.392$), ML162 ($P = 0.424$) and ML210 ($P = 0.398$). **b**, Dot plot depicting the correlation of the dependency of a cell on *GPX4* (CERES score of -1 means full dependency based on CRISPR–Cas9 knockout screen) and the expression level of *FSP1* in a panel of 559 different cancer cell lines (DepMap; <https://depmap.org/portal/>). Cell lines with high expression of *FSP1* were found to be less dependent on *GPX4* (Pearson

correlation score of 0.366 , $P = 3.38 \times 10^{-19}$). **c**, Dose-dependent toxicity of RSL3 in a panel of human lung cancer cells (NCI-H1437, NCI-H1299, NCI-H1573, NCI-H2126, NCI-H520 and NCI-H661) treated with or without the FSP1 inhibitor iFSP1 ($5 \mu\text{M}$). Co-treatment of RSL3 and iFSP1 increased the ferroptotic response of all cell lines except in NCI-H1437 cells. **d**, Dose-dependent toxicity of different cytotoxic compounds (erastin, BSO, RSL3, vinblastine, etoposide, phenylarsine oxide (PAO), mitoxantrone, irinotecan, nocodazole and cisplatin) in Pfa1 mock and FSP1-overexpressing cells treated with or without iFSP1. The protective effect of FSP1 overexpression is lost upon iFSP1 ($5 \mu\text{M}$) treatment. Data are the mean \pm s.d. of $n = 3$ wells of a 96-well plate from one representative of two independent experiments (**c**, **d**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	FlowJo software, TraceFinder software, VisiView imaging software, Image Lab 5.1 Software, Softmax Pro 6.2.1 Software, BD FACSDiva v6.1.3, cSeries Capture Software, Version 1.9.7.0802 (2017, Azure Biosystems), CorelDraw X8 Version 18.0.0.448 (2016 Corel Corporation), SparkControl V2.1 (TEcan), MultiQuant 3.0.2 Software.
Data analysis	Microsoft Excel 2016 MSO (16.0.4266.1001), Graph Pad Prism 6 (Version 6.07), Graph Pad Prism 8 (Version 8.1.0), RStudioVersion 1.1.463, Analyst® 1.7, were used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Preliminary cell viability experiments showed small variations between biological replicates, so we chose $n \geq 3$ for reproducibility. For the determination of phospholipid composition we chose $n \geq 4$ according to our experience in previous experiments (small variation between biological replicates).
Data exclusions	In very rare cases single values of biological triplicates were excluded from the analysis due to cell clumps/ uneven plating.
Replication	All attempts to replicate experiments were successful, accounting for the robustness of the results. To guarantee reliable replication of our results we pretested all used sera for their suitability for ferroptosis research. It is known that differing vitamin E and selenium concentrations in different sera batches profoundly impact on the outcome of ferroptosis inducing/inhibiting conditions.
Randomization	Not applicable as no animal studies were performed.
Blinding	Blinding was not required for cell viability measurements. For the determination of phospholipid composition the investigator performing the MS analysis and quantification was blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies against GPX4 (1:1000; no. ab125066, Abcam), ACSL4 (1:200; no. sc-271800, Santa Cruz), β -ACTIN (1:10000; no. A5441, Sigma), VCP (1:2000; ab11433, Abcam), AIFM2 (1:1000; PA5-24562, Thermo, Extended Data Fig. 1e), AIFM2 (1:10; Rat IgG2a monoclonal antibody raised against recombinant human AIFM2 protein, clone 6D8-11, developed in this study), XCT (1:10; Rat IgG2a monoclonal antibody raised against a N-terminal peptide of hXCT, clone 3A12-1-1, developed in-house), P53 (1:1000; p53 Antibody #9282, Cell Signaling), P21 (1:1000; p21 Waf1/Cip1 (12D1) Rabbit mAb #2947, Cell Signaling), cleaved caspase 3 (1:1000; cleaved caspase3 antibody (Asp175) #9661, Cell Signaling), HA tag (Rat IgG1 Anti-HA High affinity 3F10, Roche), ARL1 (1:500, no. 16012-1-AP, Proteintech) were used in this study.
Validation	Antibody against GPX4 (no. ab125066) was validated for westernblotting in a previous publication (PMID: 25402683). Antibody against ACSL4 (no. sc-271800) was validated for westernblotting in a previous publication (PMID: 27842070). Antibody against β -ACTIN (A5441) was validated as loading control for westernblotting in a previous publication (PMID 15809369). Antibody against VCP (ab11433) has been validated as loading control in westernblotting in a previous publication (PMID: 19139805). Antibody against AIFM2 (PA5-24562) has been validated for westernblotting in this study in Extended Data Fig. 1e. Antibody against AIFM2 (clone 6D8-11, developed in this study) has been validated for westernblotting in this study Extended Data Fig. 4d. Antibody against XCT (clone 3A12-1-1, developed in this study) has been validated for westernblotting in this study Fig. 4a. P53 antibody (#9282) recognizes endogenous levels of P53 stated on by cell signaling website. P21 antibody (p21 Waf1/Cip1 (12D1) Rabbit mAb #2947) specificity validated on cell signaling website. Cleaved caspase 3 antibody (Antibody #9661) specificity validated on cell signaling website.

Antibody against HA tag (clone 3F10) was validated on the manufacturer's website.
ARL1 antibody (no. 16012-1-AP) was validated on the manufacturer's website.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

4-hydroxytamoxifen (TAM)-inducible Gpx4^{-/-} murine immortalized fibroblasts (Pfa1) were reported previously (PMID: 18762024). Human fibrosarcoma (HT1080) cells, human hepatocellular carcinoma (HepG2, HLE, HLF, HUH-7, SNV182, SNV387 (HLE, HLF, HUH-7, SNV182 and SNV387 were all kind gifts from Prof. Martin Eilers, Würzburg University), human lung cancer cells (NCI-H1299, NCI-H1437, NCI-1882, NCI-H1563, NCI-H1573, NCI-H1975, NCI-H2126, NCI-H520, NCI-H661, A549, PC9, EKVX, NCI-H460), human glioblastoma cells (U-87 MG, U-251 MG, U-138 MG, U-373 MG), human neuroblastoma cells (IMR5/75 (kind gift from Dr. Frank Westermann, DKFZ Heidelberg), human breast cancer cells (MDA-MB-157, MDA-MB-231, MDA-MB-436, MDA-MB-453, MDA-MB-468, BT-474, BT-549, MCF7, T-47D, HS-578T), human colon cancer cells (SW620, SW480, RKO, LS174T, LoVo, HT29, HCT116, DLD-1), human kidney cancer cells (769-P, 786-O, LOX-IMVI, HK2, RCC4, UMRC2, UMRC3, human melanoma and uveal melanoma (M19Mel, UACC-62, LOX-IMVI, MDA-MB-435, A375, SK-Mel-2, SK-Mel-5, Mel270, OMN1-GNA11, OMN1-3 and 92-1 were all kind gifts from Prof. Svenja Meierjohann, Würzburg University) were purchased from ATCC unless stated otherwise and cultured according to ATCC guidelines. All cells were regularly tested for mycoplasma contamination.

Authentication

Non of the cell lines used were authenticated.

Mycoplasma contamination

All cell lines were tested negative for mycoplasma contamination.

Commonly misidentified lines
(See [ICLAC](#) register)

MDA-MB-435 (SAMN03151832), U-373 MG (SAMN03151977)

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Determination of apoptosis using AnnexinV/PI staining. 200,000 Pfa1 p442-Mock and Pfa1 p442-FSP1 cells were seeded on each well of a 6-well plate. On the next day, cells were treated with TNF α (10 ng/mL) for 4 h and stained according to the manufacturers protocol (eBioscience™ Annexin V Apoptosis Detection Kit FITC). Subsequently, they were analyzed on a BD FACSCANTO II instrument using the AlexaFluor 488 filter for the Fit-C labeled AnnexinV antibody and the PE-Cy5 filter for PI staining.

Assessment of lipid peroxidation using C11-BODIPY (581/591). 150,000 cells per well were seeded on 6-well dishes (Nunc) one day prior to the experiment. On the next day, cells were treated with the indicated concentration of (1S, 3R)-RSL3 to induce ferroptosis. Cells were incubated with C11-BODIPY (581/591) (1 μ M) for 30 min at 37°C before they were harvested by trypsinisation. Subsequently, cells were resuspended in 500 μ L of fresh PBS (DPBS, Gibco) strained through a 35 μ M cell strainer (Falcon tube with cell strainer CAP) and analyzed using the 488-nm laser of flow cytometer (FACS Canto II, BD Biosciences) for excitation. Data was collected from the FL1 detector (C11-BODIPY) with a 502LP and 530/30 BP filter. At least 10,000 events were analyzed per sample. Data was analyzed using FlowJo Software.

Pfa1_Cas9 cells were used to generate Pfa1 AIFM1 KO cells by lentiviral infection with ecotropic pseudotyped particles containing the desired sgRNA expressing plasmids (pKLV-U6gRNA(sgRNA)-PGKpuro2ABFP). Two days after infection, Pfa1_Cas9 cells were sorted on a BD Bioscience FACSARIA II using Blue Fluorescent Protein (BFP) as a marker.

Instrument

FACS Canto II, BD Biosciences and BD Bioscience FACSARIA II

Software

For data collection the BD FACS DIVA software was used.
For data analysis FlowJo (Version 10) software was used.

Cell population abundance

The abundance of the desired cell population in post-sort fractions was generally > 96 % of the total post sort population. Sorting was performed using the 4-way purity setting on the BD FACSARIA II.

Gating strategy

Live cell populations were separated from cellular debris and dead cells using FSC/SSC.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Circular ecDNA promotes accessible chromatin and high oncogene expression

<https://doi.org/10.1038/s41586-019-1763-5>

Received: 6 November 2018

Accepted: 26 September 2019

Published online: 20 November 2019

Sihan Wu^{1,18}, Kristen M. Turner^{1,14,18}, Nam Nguyen^{2,14,18}, Ramya Raviram¹, Marcella Erb³, Jennifer Santini³, Jens Luebeck⁴, Utkrisht Rajkumar², Yarui Diao^{1,15,16}, Bin Li¹, Wenjing Zhang¹, Nathan Jameson¹, M. Ryan Corces⁵, Jeffrey M. Granja⁵, Xingqi Chen^{5,17}, Ceyda Coruh⁶, Armen Abnoui⁷, Jack Houston¹, Zhen Ye¹, Rong Hu¹, Miao Yu¹, Hoon Kim⁸, Julie A. Law⁶, Roel G. W. Verhaak⁸, Ming Hu⁷, Frank B. Furnari¹, Howard Y. Chang^{5,9,19*}, Bing Ren^{1,10,11,19*}, Vineet Bafna^{2,19*} & Paul S. Mischel^{1,12,13,19*}

Oncogenes are commonly amplified on particles of extrachromosomal DNA (ecDNA) in cancer^{1,2}, but our understanding of the structure of ecDNA and its effect on gene regulation is limited. Here, by integrating ultrastructural imaging, long-range optical mapping and computational analysis of whole-genome sequencing, we demonstrate the structure of circular ecDNA. Pan-cancer analyses reveal that oncogenes encoded on ecDNA are among the most highly expressed genes in the transcriptome of the tumours, linking increased copy number with high transcription levels. Quantitative assessment of the chromatin state reveals that although ecDNA is packaged into chromatin with intact domain structure, it lacks higher-order compaction that is typical of chromosomes and displays significantly enhanced chromatin accessibility. Furthermore, ecDNA is shown to have a significantly greater number of ultra-long-range interactions with active chromatin, which provides insight into how the structure of circular ecDNA affects oncogene function, and connects ecDNA biology with modern cancer genomics and epigenetics.

DNA encodes information not only in its sequence, but also in its shape. The human genome is segmented into chromosomes that are made of chromatin fibres folded into dynamic, hierarchical structures^{3,4}. This spatial architecture, including numerous loops of chromatin, brings distant elements into proximity and organizes transcriptional activities into distinct compartments, restricting the accessibility of DNA to the regulatory and transcriptional machinery. In cancer, this chromatin landscape is markedly altered^{5,6}. ecDNA with amplified oncogenes was recently shown to be widespread in cancer¹, complementing the diversity of non-chromosomal DNA elements^{7,8}. ecDNA differs from the kilobase-size circular DNA found in healthy somatic tissues^{2,7,8}, because ecDNA is 100–1,000 times larger and highly amplified, raising challenging questions about ecDNA topology and how it might affect transcriptional and epigenetic regulation in cancer.

ecDNA is circular

To understand ecDNA structure, transcription and chromatin organization, we studied three human cancer cell lines (Extended Data Fig. 1a)

and clinical tumour samples from The Cancer Genome Atlas (TCGA), by integrating imaging and sequencing approaches (Fig. 1a). Whole-genome sequencing (WGS) analysis has previously been used to resolve ecDNA structure, using a computational tool—AmpliconArchitect^{1,9}—that classifies amplicons as circular or linear (Supplementary Table 1). Circular amplicons in GBM39 cells detected by this approach were confirmed to be extrachromosomal by fluorescence in situ hybridization (FISH) of tumour cells in metaphase (Fig. 1b, Extended Data Fig. 1b–d). The reconstructed circular amplicon structure was supported by many paired-end discordant junctional reads and validated by Sanger sequencing (Extended Data Fig. 1e, f). Genes detected on linear amplicons were found on chromosomal DNA (chrDNA) (Extended Data Fig. 1g). Reconstruction of 41 circular amplicons from 37 human cancer cell lines¹ revealed amplicon sizes ranging from 168 kb to 5 Mb, with a median size of 1.26 Mb (Extended Data Fig. 1h).

AmpliconArchitect infers a shape on the basis of computational reconstruction of short, paired-end reads (100–200 bp), but does not unambiguously place large duplications in the structure. To augment our understanding of ecDNA shape based on its sequence,

¹Ludwig Institute for Cancer Research, University of California at San Diego, La Jolla, CA, USA. ²Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA. ³UCSD Light Microscopy Core Facility, Department of Neurosciences, University of California at San Diego, La Jolla, CA, USA. ⁴Bioinformatics & Systems Biology Graduate Program, University of California at San Diego, La Jolla, CA, USA. ⁵Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA. ⁶Plant Molecular and Cellular Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, USA. ⁷Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA. ⁸The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. ⁹Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ¹⁰Department of Cellular and Molecular Medicine, Center for Epigenomics, University of California at San Diego, La Jolla, CA, USA. ¹¹Institute of Genomic Medicine, Moores Cancer Center, University of California at San Diego, La Jolla, CA, USA. ¹²Moores Cancer Center, University of California at San Diego, La Jolla, CA, USA. ¹³Department of Pathology, University of California at San Diego, La Jolla, CA, USA. ¹⁴Present address: Boundless Bio, Inc., La Jolla, CA, USA. ¹⁵Present address: Department of Cell Biology, Regeneration Next Initiative, Duke University School of Medicine, Durham, NC, USA. ¹⁶Present address: Department of Orthopaedic Surgery, Regeneration Next Initiative, Duke University School of Medicine, Durham, NC, USA. ¹⁷Present address: Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. ¹⁸These authors contributed equally: Sihan Wu, Kristen M. Turner, Nam Nguyen. ¹⁹These authors jointly supervised this work: Howard Y. Chang, Bing Ren, Vineet Bafna, Paul S. Mischel. *e-mail: howchang@stanford.edu; biren@ucsd.edu; vbafna@cs.ucsd.edu; pmischel@ucsd.edu

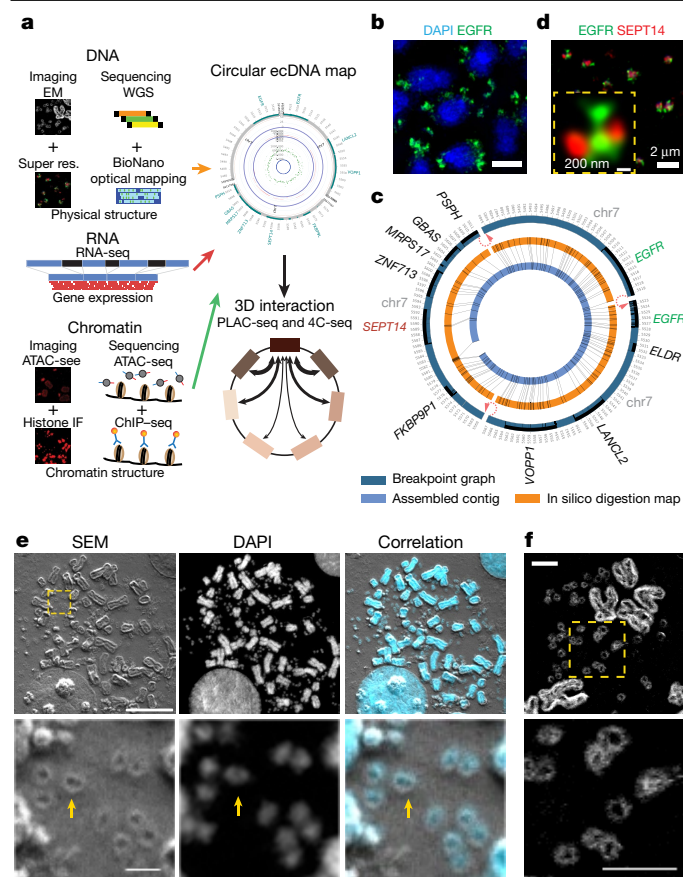


Fig. 1 | ecDNA physical structure is circular. **a**, Global workflow to characterize the structure and function of ecDNA. IF, immunofluorescence. **b**, Representative EGFR FISH in GBM39 cells. Scale bar, 5 μ m. **c**, Composite breakpoint graph generated by AmpliconArchitect, in silico digestion map and the assembled contig from BioNano optical mapping of GBM39 ecDNA. Red arrows indicate breakpoints connected by discordant paired-end WGS reads. **d**, Double FISH of EGFR and SEPT14 identified from **c**. **e**, Correlated SEM and confocal light microscopy of chromosomal and ecDNA in COLO320DM cells. Scale bars, 10 μ m (top) and 1 μ m (bottom). **f**, SEM back-scatter in COLO320DM cells. Scale bars, 2 μ m. All imaging experiments were repeated at least three times, with similar results.

we integrated optical mapping of long-range reads (approximately 160,000 bp) of DNA, using the BioNano technology platform, which permits the development of a physical map based on long contiguous pieces of DNA^{10,11}. We developed a tool, AmpliconReconstructor, to integrate the optical mapping contigs with AmpliconArchitect-based WGS reconstructions, resolving a 1.3-Mb circular, contiguous ecDNA molecule in GBM39 cells (Fig. 1c, Extended Data Fig. 2a). Individual genes on the amplicon were visualized by super-resolution confocal microscopy (Fig. 1d, Extended Data Fig. 2b).

To visualize ecDNA architecture directly, we captured images of human COLO320DM cells containing *MYC* ecDNA (Extended Data Fig. 2c) using super-resolution three-dimensional structured illumination microscopy (3D-SIM)¹², which revealed circular ecDNA particles (Extended Data Fig. 2d). To obtain more definitive evidence, we performed scanning and transmission electron microscopy (SEM and TEM). Correlative light and electron microscopy analysis of COLO320DM cells—which contain larger-size ecDNA than GBM39 cells, making them advantageous for visualization (Extended Data Fig. 1h)—demonstrated that ecDNAs stained by the fluorescent dye 4',6-diamidino-2-phenylindole (DAPI) are circular (Fig. 1e, f). TEM analysis of GBM39 cells independently confirmed the presence of circular ecDNAs, including classical double minutes^{13,14} (Extended Data Fig. 2e). Together, these results combining DNA sequencing, optical mapping,

super resolution 3D-SIM, SEM and TEM analysis demonstrate that the ecDNAs studied here are circular.

ecDNA drives massive oncogene expression

To determine the effect on transcription, we integrated RNA sequencing (RNA-seq) with WGS from cancer cell lines and from TCGA clinical tumour samples of diverse histological types, revealing that genes encoded on ecDNA—particularly bona fide oncogenes—are among the most highly expressed genes in cancer genomes (Fig. 2a, b, Extended Data Fig. 3a, b). Using our AmpliconArchitect-based approach to determine whether specific genes are amplified on circular ecDNA, we found that in cancer cell lines and clinical tumour samples, oncogenes amplified on ecDNA have markedly increased numbers of transcripts compared with the same genes when they are not amplified by circularization (Fig. 2c, d, Extended Data Fig. 3c–g). We searched for single nucleotide polymorphisms in the WGS and RNA-seq data that permitted us to distinguish between transcription from genes on ecDNA and their native chromosomal loci, revealing massively increased transcription from genes encoded on ecDNAs (Fig. 2e). In fact, oncogenes encoded on ecDNA—including *EGFR*, *MYC*, *CDK4* and *MDM2*—are among the top 1% of genes expressed in the cancer genomes (Fig. 2b, Supplementary Table 2).

The amount of RNA transcribed can be related to the amount of available DNA template. We hypothesized that the massively increased oncogene transcription on ecDNA is likely to be driven by their increased DNA copy number¹⁵ (Extended Data Fig. 3g, h). Accordingly, oncogenes amplified on ecDNA were shown to achieve far higher copy numbers than the same genes amplified on linear structures (Fig. 2f, g). However, the amount of DNA template is not the only factor that determines gene transcription. Chromatin organization influences the accessibility of DNA to the regulatory machinery of transcription^{4,16}. In some cases, oncogenes on ecDNA produced more transcripts, even when normalized to gene copy number (Extended Data Fig. 3g, h). We initiated a deeper examination of other chromatin structural features that may contribute to the massively increased expression of oncogenes amplified on ecDNA.

ecDNA contains highly accessible chromatin

Most of the human genome is not transcribed in a given cell because it is tightly wound around histone octamers that in turn are packed into complex hierarchical structures, rendering the DNA inaccessible to transcription factors and the transcription machinery^{17,18}. We used complementary approaches to resolve the ecDNA chromatin landscape. First, we analysed active and repressive histone marks by immunofluorescence analysis of cancer cells in metaphase and also performed H3K4me1 and H3K27ac chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) analyses of actively cycling GBM39 cells, which revealed the presence of active histone marks on ecDNA¹⁹ (Extended Data Fig. 4a–c), and a concomitant paucity of repressive histone mark on GBM39 ecDNA (Extended Data Fig. 4d, e). Second, we used the assay for transposase-accessible chromatin using sequencing (ATAC-seq) and micrococcal nuclease digestion and sequencing (MNase-seq) to assess chromatin accessibility and to map nucleosome positions. Finally, we used the assay of transposase-accessible chromatin with visualization (ATAC-seq) to visualize accessible chromatin directly²⁰ (Extended Data Fig. 5a). The periodic length distributions of DNA fragments generated by ATAC-seq and MNase-seq demonstrated that ecDNA is packaged into chromatin, and consists of nucleosome units (Fig. 3a, Extended Data Fig. 5b, c). However, ecDNA displayed a significant deficit in the number of long fragments (more than 1,200 bp) from ATAC-seq and MNase-seq, indicative of compacted nucleosomal arrays (Fig. 3a, Extended Data Fig. 5b, c), and a significantly increased number of ATAC-seq peaks (Fig. 3b,

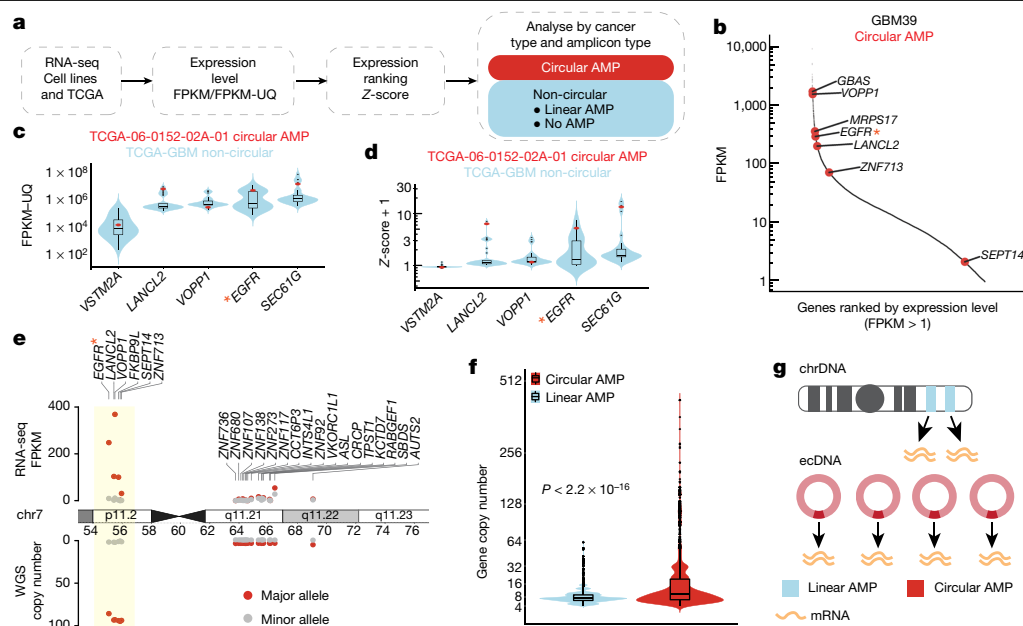


Fig. 2 | ecDNA drives high levels of RNA expression. **a**, Workflow of RNA-seq data analysis. FPKM, fragments per kilobase of transcript per million mapped reads; FPKM-UQ, FPKM upper quartile. **b**, ecDNA gene expression within the transcriptome of GBM39 cells. Red dots denote genes on ecDNA (circular amplification). *GBAS* is also known as *NIPSNAP2*; *SEPT14* is also known as *SEPTIN14*. **c**, ecDNA gene expression in one TCGA-GBM sample (red data points) compared to non-circular genes in the TCGA-GBM cohort (blue violin and box plot distribution) ($n = 36$ biologically independent samples). **d**, Z-score of the gene expression plotted in **c**. Z-scores were plotted as +1 to avoid negative

values during \log_{10} transformation. **e**, Allele-specific gene copy number and mRNA expression levels in GBM39 cells. Circular amplified region (ecDNA) was highlighted. **f**, Gene copy number comparing circular and linear amplifications (8,068 circular and 6,247 linear amplified (AMP) genes from 77 samples). P value determined by two-sided Wilcoxon test. **g**, Depiction of the mechanism of massive transcript levels from ecDNA. Asterisks in **b–e** indicate key oncogenes. Violin plots show the overall distribution of data points. Box plots show median, upper and lower quartiles; whiskers indicate 1.5 times the interquartile range, and black points are the outliers.

Extended Data Fig. 5d), which suggests that the ecDNA chromatin landscape is more accessible than chrDNA, because its nucleosomal organization is less compacted.

The recent landmark study deciphering the chromatin accessibility landscape in primary cancer samples⁵ enabled us to examine chromatin accessibility in authentic clinical samples. By integrating ATAC-seq profiles with WGS data analysed by AmpliconArchitect, we found a significantly higher ATAC-seq signal in the DNA with predicted circular amplicons in clinical tumour samples, even after normalizing for DNA copy number (Fig. 3c, Extended Data Fig. 5e). Even in isogenic cell lines, ecDNA is more accessible than the same locus amplified as a homogeneous staining region (HSR)²¹ on chromosomes (Extended Data Fig. 5f–h). Notably, the HSR region did not show a deficit in the number of long ATAC-seq fragments as compared to ecDNA (Extended Data Fig. 5i). We further validated that both the enhanced chromatin accessibility and active chromatin states are linked to the increased rates of transcription from the allele contained on highly amplified ecDNA (Extended Data Fig. 5j).

We then applied the ATAC-seq technology to analyse accessible chromatin in actively cycling cells in interphase by staining COLO320DM cells with ATAC-seq and DAPI to label accessible chromatin and DNA, respectively, and to permit the sorting of tumour cells in early G1 phase²⁰, followed by MYC fluorescence in situ hybridization (FISH) to label ecDNAs. A notable positive correlation between the ecDNA-containing MYC FISH signal and the ATAC-seq signal was seen, which demonstrates highly accessible chromatin of ecDNA at single-cell resolution (Fig. 3d, e, Extended Data Fig. 6a–c). ecDNA remained similarly accessible during metaphase (Extended Data Fig. 7a–d). Together, these data demonstrate that some of the most accessible chromatin in the genome of cancer cells resides on ecDNA, possibly owing to the lower level of chromatin compaction (Fig. 3f). In fact, ATAC-seq enabled us to identify unanticipated MYC ecDNAs in GBM39 cells because of their high signal, which was subsequently

confirmed by ATAC-seq and WGS (Extended Data Fig. 7c, Supplementary Table 1).

To contextualize these genetic, transcriptional and epigenetic features, we generated circular maps of ecDNA in cancer cell lines and primary tumour samples (Fig. 4a, Extended Data Fig. 8). These topologically informed maps highlighted the high DNA copy number, high levels of transcription particularly of its constituent oncogenes, and high accessibility of its chromatin, bridging ecDNA circular structure with biological function. ecDNAs within a tumour can also vary in size and composition (that is, sequence), even when they contain the same oncogene. In GBM39 cells, the structures of EGFR-containing ecDNAs are uniform (Extended Data Fig. 9). Consequently, the WGS trace in its circular map is relatively uniform (Fig. 4a). By contrast, COLO320DM and PC3 cells contain diverse MYC-containing ecDNA populations, which results in a more heterogeneous WGS trace in the circular ecDNA plots (Extended Data Figs. 8a, b and 9).

ecDNA enables ultra-long-range chromatin contacts

We performed proximity ligation-assisted ChIP-seq²² (PLAC-seq, similar to HiChIP²³) to map the 3D chromatin interactions genome-wide anchored at DNA bound by histone with H3K27ac modification in GBM39 cells. We also conducted circular chromosome conformation capture combined with high-throughput sequencing (4C-seq) to provide an independent assessment of chromatin contacts in GBM39 cells. Together with ChIP-seq of CTCF and cohesin subunit protein SMC3 to examine the locations of factors that are important for the organization of chromatin domains²⁴, these data revealed a massive increase in diagonal corner reads in the GBM39 ecDNA junctional region (Fig. 4b, Extended Data Fig. 10a), and rebound of the virtual 4C signal in the distal region (Extended Data Fig. 10b), providing further orthogonal evidence that ecDNA is circular (Fig. 4c). In addition, the binding of CTCF and cohesin demonstrate that ecDNA chromatin

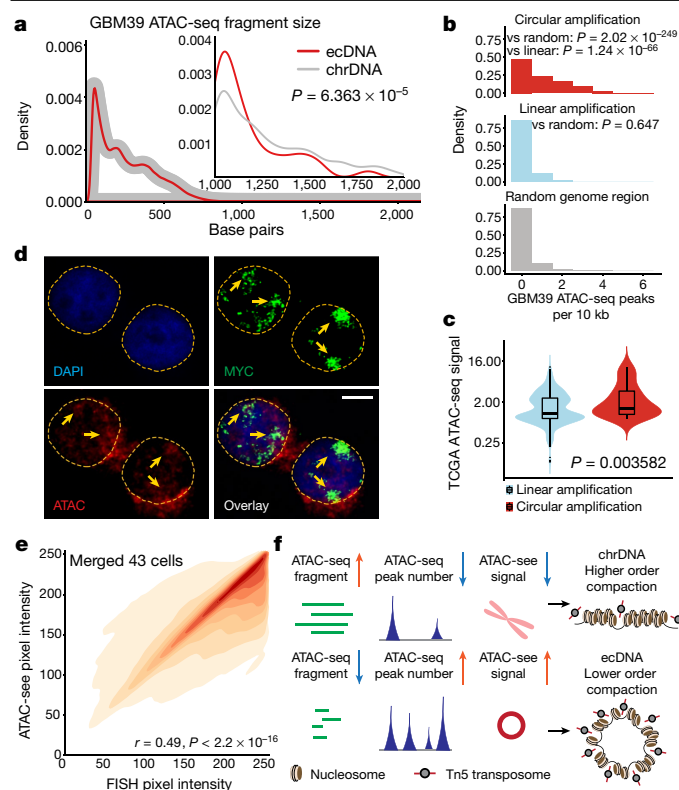


Fig. 3 | The chromatin landscape of ecDNA. **a**, Global and long (>1 kb) ATAC-seq fragment size distribution of ecDNA and chrDNA (110 ecDNA and 1,571 chrDNA long fragments; $n = 2$ biologically independent samples, showing one of the representative results). P value determined by two-sided Kolmogorov–Smirnov test. **b**, ATAC-seq peak number per 10 kb in GBM39 cells (circular, 714 windows; linear, 268 windows; random, 313,762 windows; $n = 2$ biologically independent samples). P values determined by Kruskal–Wallis test. **c**, TCGA ATAC-seq read counts normalized by copy number (circular: 8 samples, 33 amplicons; linear: 7 samples, 476 amplicons). Violin plots show the overall distribution of data points. Box plots as in Fig. 2g. P value determined by Z-test. **d**, Co-localization of the ATAC-seq and FISH signal in interphase cell nuclei from COLO320DM cells. Scale bar, 5 μ m. **e**, Pearson correlation of FISH and ATAC-seq signal pixel intensity, merged from 43 COLO320DM single cells in interphase. **f**, Depiction and interpretation of integrated technologies to assess chromatin compaction.

independently demonstrated ultra-long-range chromatin contacts that can occur on ecDNA (Extended Data Fig. 10c, d), which could potentially have some effect on distal gene expression, as suggested by CRISPR interference targeting catalytically inactive Cas9 (dCas9) fused to the KRAB transcriptional repressor domain to mask the *EGFR* promoter (Extended Data Fig. 10e–j).

Amplification of oncogenes on ecDNA is surprisingly prevalent in cancer^{1,25}, and it can markedly increase oncogene copy number and drive intratumoural genetic heterogeneity because it lacks centromeres and is subject to unequal segregation^{1,26}. These results demonstrate that ecDNA promotes massively increased transcription of the oncogenes studied here, owing to its increased DNA copy numbers and in association with enhanced chromatin accessibility, highlighting a mechanism by which ecDNA contributes to cancer pathogenesis by altering the shape of its chromatin.

In bacteria, small circular plasmids represent a prevalent and powerful mechanism for rapidly gaining selective advantage²⁷. We speculate that oncogene-containing circular ecDNA in human cancers represents the conceptual equivalent, highlighting crucial gene variants and mechanisms for oncogenesis and therapeutic resistance^{28–30}.

is well organized, indicative of topologically associating domains (Fig. 4b). Furthermore, downsampling the PLAC-seq reads from the GBM39 ecDNA region to a level comparable to the same region in U87 cells that lack ecDNA demonstrated notably increased distal interactions in active chromatin on ecDNA (Fig. 4c, Extended Data Fig. 10a, b). Using the *EGFR* promoter as bait, the virtual 4C and actual 4C-seq

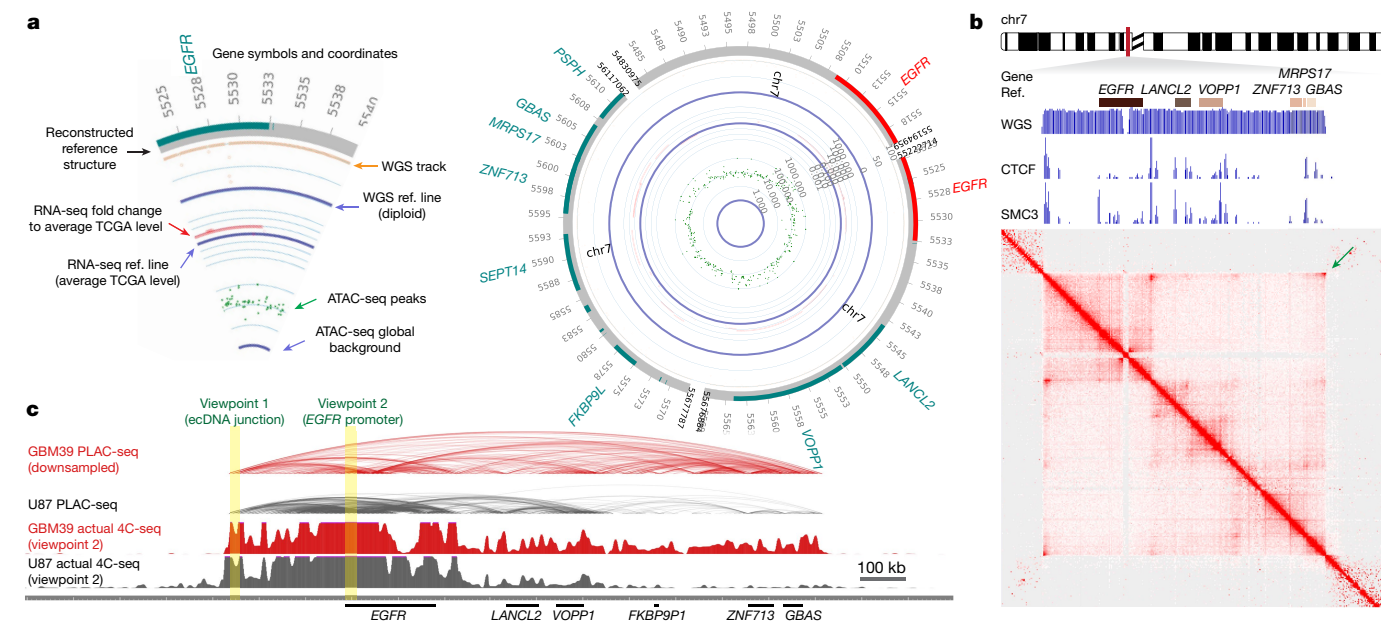


Fig. 4 | Circularization of ecDNA enables distal DNA interaction. **a**, Circular plot of ecDNA structure in GBM39 cells. Plot legend is shown on the left. **b**, MAGIC Collaboration H3K27ac anchored active chromatin interaction heatmap by PLAC-seq and HiChIP in GBM39 cells. WGS depicts the ecDNA amplicon.

ChIP-seq demonstrates CTCF and SMC3 binding to ecDNA. Arrow indicates the increased corner reads in ecDNA junction. **c**, Composite view of PLAC-seq, HiChIP and actual 4C-seq. Virtual and actual 4C-seq viewpoints are highlighted.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1763-5>.

1. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
2. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer* **19**, 283–288 (2019).
3. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).
4. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
5. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
6. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
7. Möller, H. D. et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nat. Commun.* **9**, 1069 (2018).
8. Shibata, Y. et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. *Science* **336**, 82–86 (2012).
9. Deshpande, V. et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 392 (2019).
10. Mendelowitz, L. & Pop, M. Computational methods for optical mapping. *Gigascience* **3**, 33 (2014).
11. Mak, A. C. et al. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**, 351–362 (2016).
12. Demmerle, J. et al. Strategic and practical guidelines for successful structured illumination microscopy. *Nat. Protocols* **12**, 988–1010 (2017).
13. Schimke, R. T. Gene amplification in cultured animal cells. *Cell* **37**, 705–713 (1984).
14. Storlazzi, C. T. et al. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res.* **20**, 1198–1206 (2010).
15. L'Abbate, A. et al. MYC-containing amplicons in acute myeloid leukemia: genomic structures, evolution, and transcriptional consequences. *Leukemia* **32**, 2152–2166 (2018).
16. Baylin, S. B. & Jones, P. A. Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* **8**, a019505 (2016).
17. Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).
18. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
19. Smith, G. et al. c-Myc-induced extrachromosomal elements carry active chromatin. *Neoplasia* **5**, 110–120 (2003).
20. Chen, X. et al. ATAC-seq reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods* **13**, 1013–1020 (2016).
21. Solovei, I. et al. Topology of double minutes (dmins) and homogeneously staining regions (HSRs) in nuclei of human neuroblastoma cell lines. *Genes Chromosom. Cancer* **29**, 297–308 (2000).
22. Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
23. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
24. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
25. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e318 (2018).
26. deCarvalho, A. C. et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).
27. Lederberg, J. Cell genetics and hereditary symbiosis. *Physiol. Rev.* **32**, 403–430 (1952).
28. Nathanson, D. A. et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* **343**, 72–76 (2014).
29. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
30. Xu, K. et al. Structure and evolution of double minutes in diagnosis and relapse brain tumors. *Acta Neuropathol.* **137**, 123–137 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

Cell culture

Human prostate cancer cell line PC3, colon cancer cell line COLO320DM and glioblastoma cell line U87 were purchased from ATCC and cultured in DMEM/F12 with 10% fetal bovine serum (FBS). Human glioblastoma GBM39 tumour spheroid was derived from patient tissue, and cultured in DMEM/F12 with GlutaMAX, B27, 20 ng ml⁻¹ EGF, 20 ng ml⁻¹ FGF and 5 µg ml⁻¹ heparin. All cell lines tested negative for mycoplasma.

Metaphase chromosome spread

Cells in metaphase were obtained by KaryoMAX (Gibco) treatment at 0.1 µg ml⁻¹ for 3 h (PC3 and COLO320DM) or overnight (GBM39). Cells were washed with PBS and single cells were suspended in 75 mM KCl for 15–30 min. Samples were then fixed by Carnoy's fixative (3:1 methanol:glacial acetic acid, v/v) and washed an additional three times with fixative before being dropped onto humidified glass coverslips.

FISH

Coverslips containing fixed cells in metaphase were aged overnight, briefly equilibrated by submerging in 2× SSC buffer, followed by dehydration in ascending ethanol series (70%, 85%, 100%) for 2 min each. Pre-warmed FISH probes (Empire Genomics) were added onto a slide, and the coverslip was applied and sealed with rubber cement. The FISH probe and sample were co-denatured on a 75 °C hotplate for 3 min, and hybridization was carried out overnight at 37 °C in a humidified chamber. The coverslips were removed and washed in 0.4× SSC at 72 °C, followed by a final wash in 2× SSC, 0.05% Tween-20, for 2 min each. DNA was stained with DAPI (1 µg ml⁻¹; 2 min), washed with 2× SSC, and then mounting medium (VectaShield) was applied and the coverslip was mounted onto a glass slide.

Immunofluorescence on metaphase chromosome

Metaphase cells were obtained similarly by KaryoMAX treatment and KCl swelling. Unfixed cells (2.5×10^4 – 4×10^4) were spread onto a slide by Cytospin cytocentrifuge (Thermo Scientific). After ageing overnight at 4 °C, 100 µl primary and secondary antibodies in antibody diluent (DAKO) were applied sequentially onto the samples, with gentle washing by 2× SSC buffer with 0.1% Tween-20. Samples were then fixed by 4% paraformaldehyde in PBS, rinsed and mounted with ProLong Gold antifade mounting medium with DAPI (Invitrogen). The primary antibodies were: anti-H3K4me1 (CST 5326), anti-H3K27ac (CST 8173), anti-H3K4me3 (Diagenode C15410003), anti-H3K18ac (Diagenode C15410139), anti-H3K9me3 (Active Motif 39765), and anti-H3K27me3 (Active Motif 39155).

Correlative light and electron microscopy

Fixed cells in metaphase were dropped onto Zeiss coverslips with fiducial markings. Images of DAPI-stained cells were captured with a Zeiss 880 Airyscan confocal microscope, and the locations of select cells were stored using the Shuttle and Find feature of the ZEN Black software. To correlate SEM with the DAPI-stained acquired images, the coverslip was briefly washed with ddH₂O and stained with 2% uranyl acetate for 2 min. The coverslip and holder were then loaded into the SEM and the same previously imaged DAPI-stained cells in metaphase were located using Shuttle and Find with ZEN Blue software. Images were captured using a Zeiss Sigma VP Scanning Electron Microscope and correlated with light microscope images.

Structured illumination microscopy

Cells in metaphase were prepared and dropped onto a glass coverslip. FISH was carried out as described, and images were captured with a GE (formerly Applied Precision) DeltaVision OMX V2 Structured Illumination microscope with a 100× Olympus PlanApo 1.4 NA objective and EMCCD 10 MHz camera mode. Structured illumination reconstructions

were performed using Softworx v.6.5.2, with the Wiener filter for 442 channel set to 0.0060. Volume renderings were also done with Softworx v.6.5.2 software via the RGB opacity method, and these were used to generate 3D intensity plots of ecDNA.

TEM

Cells in metaphase were dropped onto a glass coverslip and fixed in 2% glutaraldehyde, 0.1 M cacodylate buffer. The sample was then stained in a 1% osmium tetroxide in 0.15 M cacodylate buffer for 1 h on ice, followed by three washes in 0.1 M cacodylate buffer for 15 min each. Cells were then immersed in 2% uranyl acetate in water for 1 h on ice and dehydrated in a graded series of ethanol (20%, 50%, 70%, 90% and 100%) on ice for 15 min each. The sample was then embedded in Durcupan resin and polymerized overnight in a 60 °C oven, sectioned at 50–60 nm on a Leica UCT7 ultramicrotome, and picked up on a Formvar and carbon-coated copper grid. Sections were post-stained with 2% uranyl acetate for 5 min and Sato's lead stain for 1 min. Images were captured at 25 kX using a Jeol 1400Plus TEM equipped with a 16 megapixel Gatan OneView camera.

Confocal microscopy

Immunofluorescence and ATAC-seq images were acquired by Zeiss LSM880 Airyscan confocal microscope, using 63× Plan-APOchromat NA 1.4 oil lens. Approximately 20–30 Z-stacks (4.78-µm depth) were taken from each visual field, and Fast Airyscan processing was done by ZEN Black software in 3D mode at default settings (Wiener filter was 3.3, 3.9 and 4.2 for ATAC-seq (red), MYC FISH (green), and DAPI (blue), respectively). Representative images were selected from the Z-stack with best brightness. The gain was 745, 785 and 700 for the red, green and DAPI channels, respectively. The pinhole was automatically opened by the software for Fast Airyscan acquisition, and the pixel dwell time was 0.93 µs with no averaging. Double FISH images were captured with the Leica TCS SP8 confocal microscope. Image processing for highest resolution were obtained using the Leica Lightning Imaging Information Extraction Software. We used the proprietary 'adaptive' algorithm included in the Lightning software, including the following parameters: the pinhole was set to 0.5 airy units, with no cut-off, and 4 iterations were obtained per channel. The effective resolution achieved was 118 nm and was calculated using the half-width at half-maximum method, and measured from a single FISH signal.

WGS

Genomic DNA was extracted from cells using Qiagen kits. Sequencing libraries were prepared using TruSeq adapters (Illumina) and the KAPA HyperPlus kit, according to manufacturer's instructions (Kapa Biosystems). In brief, 250 ng of DNA was used as input and enzyme-fragmented for 12 min to obtain mode fragment lengths of 350 bp. KAPA pure beads were used for double-sided size selection of 250–450 bp. DNA libraries were pooled and paired-end DNA sequencing (150 cycles) was performed on the NovaSeq S4.

AmpliconArchitect

After the FASTQ files were aligned to the reference genome using bwa mem with default parameters, AmpliconArchitect was run on the aligned reads using all regions with copy number greater than five as seeds. Default parameters were used as described in the documentation (<https://github.com/virajbdeshpande/AmpliconArchitect>). Given mapped reads, AmpliconArchitect automatically searches for other intervals participating in the amplicon, and then uses a carefully calibrated combination of copy number variant (CNV) analysis and structural variant analysis. AmpliconArchitect uses structural variant signatures (for example, discordant paired-end reads and CNV boundaries) to partition all intervals into segments and build an amplicon graph. It assigns copy numbers to the segments by optimizing a balanced flow on the graph. As short reads do not span long repeated

segments, they cannot disambiguate between multiple alternative structures. Therefore, high-molecular-mass DNA was used to generate optical mapping reads. The optical map reads were used to scaffold and disambiguate the graph, as described below.

Gene classification

To predict putative ecDNA structures, a depth-first search algorithm was used to traverse the amplicon graph and identify cycles. Genes that lay on any cycle in the graph were designated as circular. Otherwise, they were designated as linear.

Isolation of high-molecular-mass DNA for optical mapping

High-molecular-mass DNA was extracted from GBM39 cells following manufacturer's instructions (BioNano Genomics 30026) with some modifications. The initial step in the procedure calls for the generation of agarose plugs containing the cell equivalent of approximately 3–9 μg of DNA (approximately 0.5–1.5 million diploid human cells), which is a crucial step for recovering good-quality high-molecular-mass DNA. As GBM39 cells contain a roughly tetraploid amount of DNA with numerous extrachromosomal DNA¹, optimization of the DNA concentration was carried out as follows. Approximately 4.5 million GBM39 cells were spun down at 300g for 10 min, washed twice with 0.5 ml cold cell buffer (BioNano 30026), and resuspended in 450 μl cold cell buffer. This solution was then split into three different tubes to approximate 9 μg of DNA (about 0.75 million cells), 6 μg of DNA (about 0.5 million cells), or 3 μg of DNA (about 0.25 million cells), spun down at 300g for 5 min and resuspended in cell buffer to reach a final volume of 66 μl . Then, 40 μl of 2% agarose (BioRad CleanCut Agarose 170-3594) was added to the cells and incubated at 4 °C for 15 min to generate the agarose plugs. Within the plugs, the cells were lysed and digested with Proteinase K (Puregene 158920) and RNase A (Puregene 158922) per manufacturer's instructions. To stabilize, recover and clean the DNA, plugs were treated according to the manufacturer's instructions (BioNano Genomics 30026). After dialysis, the DNA was homogenized and mechanically sheared by slowly pipetting the entire volume up and down with a non-filtered 200 μl tip until the sample reached an even consistency. The DNA was then equilibrated at room temperature for 3 days. Using a 2- μl aliquot, the DNA was diluted in Qubit BR buffer, sonicated for 10 min, and quantified using the Qubit dsDNA BR Assay kit (Invitrogen Q32850). The sample obtained from the plug with around 0.5 million cells yielded the best results with a mean DNA concentration of 61 $\text{ng } \mu\text{l}^{-1}$ and a coefficient variation of 6.7% and was used for the nicking, labelling, repairing and staining reactions.

Optimization of the labelling, repairing and staining reactions and DNA loading onto IrysChip

The 2 \times nicking reaction (using Nt.BspQI) and 1 \times labelling, repairing and staining reactions were performed as per manufacturer's instructions (BioNano Genomics 30024) using the recommended NEB reagents. Using a 2- μl aliquot, the DNA was sonicated for 20 min and the final DNA concentration was determined to be 3 $\text{ng } \mu\text{l}^{-1}$ by Qubit dsDNA HS Assay kit (Life Technologies Q32854). A total of 16 μl of nicked, labelled, repaired and stained DNA was loaded onto the IrysChip (BioNano FC-020-01) and run conditions were optimized on the Irys system to ensure efficient DNA loading onto the nanochannels using the Irys User Guide (BioNano Genomics 30047).

BioNano data analysis

Thirteen rounds of data (each round containing 30 cycles of data generation) were collected on the Irys platform to reach 0.791 \times reference coverage with molecules. Raw images were processed, and long DNA molecules were detected and digitized by BioNano image-processing and analysis software AutoDetect³¹. Optical maps were generated by transforming the raw images into raw BNX files using

the IrysView software system. The BNX files output from the BioNano instrument were then assembled into optical map contigs using the BioNano Irys assembly pipeline (v.5122, default parameters). The segments discovered by Amplicon Architect were converted to an in-silico CMAP reference file and it was aligned to the assembled optical map contigs using AmpliconReconstructor (<https://github.com/jluebeck/AmpliconReconstructor>). Alignment results were also confirmed using the BioNano RefAligner (v.5122, default parameters). We produced a visualization of the resulting alignment using CycleViz (<https://github.com/jluebeck/CycleViz>).

RNA-seq

One microgram of RNA extracted by RNeasy mini kit (Qiagen) was prepared for sequencing with TruSeq RNA Library Prep Kit v2 (Illumina) according to the manufacturer's instruction. In brief, after poly-A selection and fragmentation of the total RNA, first- and second-strand cDNA was synthesized and ligated with sequencing adaptor. Products were then amplified for paired-end sequencing. Data were processed following the TCGA mRNA analysis pipeline. Expression level of mRNA was computed as FPKM for cell line samples, or as FPKM-UQ for both cell line and TCGA samples. The Z-score for FPKM-UQ was calculated as $Z\text{-score} = (X - \mu) / \sigma$, in which X is the FPKM-UQ of a given gene, μ and σ are the global mean and standard deviation, respectively, of FPKM-UQ of a given sample's transcriptome.

ChIP-seq

Formaldehyde-crosslinked chromatin from 5 million cells per ChIP was sheared to small fragments by Covaris M220 Focused-ultrasonicator. The following antibodies were used for chromatin pull-down in RIPA buffer with protease inhibitor cocktail: anti-H3K27ac (Active Motif 39685, 5 μg), anti-H3K4me1 (Active Motif 39297, 10 μl), anti-CTCF (Abcam ab70303, 5 μg), anti-SMC3 (Abcam ab9263, 5 μg). After capturing by Protein A/G magnetic beads, chromatin was washed 6 times and reverse-crosslinked for 3 h with RNase A and proteinase K. The ChIP DNA library was constructed by NEBNext Ultra II DNA Library Prep kit for paired-end sequencing.

MNase-seq

One million cells were washed by calcium-free PBS and resuspended in 1 ml lysis buffer (10 mM pH 7.5 Tris-HCl, 10 mM NaCl, 3 mM MgCl₂, 0.5% IGEPAL CA-630, 0.15 mM spermine, 0.5 mM spermidine, with Roche EDTA-free complete protease inhibitor cocktail) on ice for 5 min. After centrifugation, cell pellets were resuspended in 160 μl digestion buffer (10 mM pH 7.5 Tris-HCl, 15 mM NaCl, 60 mM KCl, 0.15 mM spermine, 0.5 mM spermidine, with protease inhibitor cocktail) on ice. Then, 0.004 U of micrococcal nuclease (NEB) in 40 μl digestion buffer (with 5 mM CaCl₂) was added to the suspension and incubated at room temperature for 10 min. Digestion was halted by 200 μl stop buffer (20 mM EDTA, 20 mM EGTA, 1% SDS). DNA was then extracted, repaired by Fast DNA End Repair Kit (Thermo Scientific), adenylated by Klenow fragment (NEB), ligated with TruSeq adapters (Illumina) and amplified to make paired-end sequencing library.

ATAC-seq

The protocol was adapted from a previous report³². In brief, 100,000–500,000 cell nuclei were extracted by NPB buffer (5% BSA, 0.2% IGEPAL-CA630, 1 mM DTT, EDTA-free protease, in PBS) at 4 °C for 10 min. Tagmentation was done in TB buffer (33 mM Tris-acetate pH 7.8, 66 mM K-acetate, 11 mM Mg-acetate, 16% DMF) with Tn5 transposase (Illumina), at 37 °C for 30 min. DNA samples were then extracted and DNA libraries were generated by PCR. To compare ATAC-seq signal between circular and linear amplicons of TCGA samples, the normalized read counts were further normalized by segment length, DNA copy number and the normalized read counts of the same length from a set of merged normal tissue controls.

Article

ATAC-seq

ATAC-seq on interphase cells was performed as previously described²⁰ and applied FISH afterward. We sorted the low ATAC-seq signal population in G1 phase by flow cytometry for confocal imaging. Protocol was modified to apply ATAC-seq on metaphase chromosome spreads. In brief, metaphase sample was prepared as described onto a 1-mm coverslip and incubated with 50 nM of ATTO-590 transposome under 37 °C for 30 min in the dark. After washed twice by 2× SSC with 0.01% SDS for 15 min, and once by 2× SSC with 0.2% Tween-20 for 15 min, sample was subjected to FISH procedures, and finally stained by 1 μg ml⁻¹ DAPI and mount with VECTASHIELD antifade mounting media (Vector Laboratories).

ATAC-seq image analysis pipeline

For ATAC-seq on interphase cell images, ImageJ was used to generate the surface plot for each colour channel, to document the pixel intensity and XY coordinate for Pearson's correlation analysis. For ATAC-seq on metaphase chromosome spreads, the ECdetect software¹ was further developed to analyse high-resolution images and semantically segment DAPI-stained nuclei, chromosomes and ecDNA. For each image, the ATAC-seq intensity at each pixel location was captured by reading the pixel values. The pixel values were then grouped based on whether they belong to ecDNA, chromosomes, or nuclei, based on the semantic segmentation information from ECdetect. This was done by comparing the pixel locations of the ATAC-seq intensities with the pixel locations of the segmentations.

PLAC-seq

Long-range chromatin interaction was probed by PLAC-seq as previously described^{22,23} using H3K27ac as the anchor (Diagenode C15200184-50), and applied MAPS pipeline³³ for the downstream data analysis. After removing PCR duplicates from the valid mapped reads, we kept all intrachromosomal reads >1 kb to quantify protein-mediated long-range chromatin interactions, and all intrachromosomal reads ≤1 kb on different strands to quantify ChIP enrichment level. Finally, we merged two replicates of the same cell type, resulting in approximately 240 million and 218 million paired-end reads for GMB39 and U87 cells, respectively. To visualize chromatin interaction frequency at the *EGFR* locus, we first selected all paired-end reads within the 1.3-Mb region (chr7: 54,830,975–56,117,062), and removed any reads overlapped with two deletion regions (chr7: 55,194,960–55,222,713, chr7: 55,676,885–55,677,786) in GBM39 cells. Because this region is highly amplified as ecDNA in GBM39 cells, resulting in much more reads, we downsampled reads in GBM39 sample to match the total number of reads at the same locus in U87 cells. Virtual 4C was generated at 10-kb resolution.

4C-seq

Five million cells were cross-linked with 2% formaldehyde for 10 min at room temperature and quenched by 125 mM glycine for 5 min. Nuclei were isolated and digested with Csp6I (Thermo Scientific) overnight. Enzyme was inactivated by heating at 65 °C for 20 min and the digested chromatin was subjected for ligation by T4 ligase (Life Technologies) for 16 h. DNA was then purified with before the second digestion with DpnII (NEB) overnight. After enzyme inactivation, a second round of ligation was performed, and DNA was purified. Then, 4.8 μg of DNA in total was used for PCR amplification (primer information in Supplementary Table 3). The 4C-seq data were analysed using 4C-ker³⁴. Reads were mapped to a reduced genome of unique 22-bp sequences flanking Csp6I sites in the hg19 genome.

CRISPR interference

Small guide RNAs (sgRNAs) targeting the *EGFR* promoter within the 4C viewpoint were cloned into pLV-hU6-sgRNA-hUbc-dCas9-KRAB-T2a-Puro (Addgene plasmid 71236)³⁵ and lentiviruses were produced by

transfecting 293T cells (sgRNA sequences in Supplementary Table 3). GBM39 cells were then infected by lentivirus (multiplicity of infection of 3) for 4 days and subjected to RNA extraction and quantitative PCR (qPCR) (qPCR primers in Supplementary Table 4).

Immunoblotting

After transferring whole-cell lysates to nitrocellulose membrane, the following antibodies were applied: anti-EGFR at 1:5,000 (EMD Millipore 06-847), anti-phospho-EGFR at 1:1,000 (CST 3777S), anti-tubulin at 1:2,000 (CST 2125S), and secondary anti-rabbit IgG antibody at 1:2,000 (CST 7074S).

Statistics

All sample sizes and statistical methods were indicated in the corresponding figure legends. No statistical methods were used to predetermine sample size. If the data were normally distributed (by Shapiro–Wilk test) and homoscedastic (by Bartlett's test), Student's *t*-test (for two groups) and one-way analysis of variance (ANOVA) (more than two groups) were used to test the mean difference. Otherwise, Wilcoxon rank-sum test (for two groups) and Kruskal–Wallis rank-sum test (for more than two groups) were applied. For ATAC-seq long fragment size distribution data, a Kolmogorov–Smirnov test was used. For the ATAC-seq signal intensity dataset, which has at least 3,500 pixels sampled for ecDNA or chrDNA per image, a Z-test was used to test the mean difference according to the central limit theorem. All statistical tests are two-sided. All box plots are shown with median, upper and lower quartiles; whiskers indicate 1.5× interquartile range, and points as outliers. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

WGS, RNA-seq, ATAC-seq, MNase-seq, ChIP-seq and PLAC-seq data are deposited in the NCBI Sequence Read Archive, under BioProject accession PRJNA506071. Source Data for Figs. 2, 3 and Extended Data Figs. 1–6, 10 are provided with the paper. Source data of the pixel quantification of ATAC-seq on metaphase chromosome spread images in Extended Data Fig. 7d are available on Figshare (<https://doi.org/10.6084/m9.figshare.9826115.v1>).

Code availability

The following are available for use online: AmpliconArchitect (<https://github.com/virajbdeshpande/AmpliconArchitect>), AmpliconReconstructor (<https://github.com/jluebeck/AmpliconReconstructor>), and CycleViz (<https://github.com/jluebeck/CycleViz>)

31. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).
32. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
33. Juric, I. et al. MAPS: model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLOS Comput. Biol.* **15**, e1006982 (2019).
34. Raviram, R. et al. 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-seq experiments. *PLOS Comput. Biol.* **12**, e1004780 (2016).
35. Thakore, P. I. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).

Acknowledgements We thank members of the Mischel laboratory, M. Farquhar for the use of the UCSD/CMM electron microscopy facility, T. Merloo and Y. Jones for electron microscopy sample preparation, UCSD Neuroscience Microscopy Shared Facility (NS047101) for providing imaging support, and the Ecker laboratory at the Salk Institute for Biological Studies for use of the Irys instrument for BioNano optical mapping. This work was supported by the Ludwig Institute for Cancer Research (P.S.M., B.R., F.B.F.), Defeat GBM Program of the National Brain

Tumor Society (P.S.M., F.B.F.), NVIDIA Foundation, Compute for the Cure (P.S.M.), The Ben and Catherine Ivy Foundation (P.S.M.), and Ruth L. Kirschstein National Research Service Award NIH/NCI T32 CA009523 (R.R.). This work was also supported by the following National Institutes of Health (NIH) grants: NS73831 (P.S.M.), R35CA209919 (H.Y.C.), RM1-HG007735 (H.Y.C.), GM114362 (V.B.), NS80939 (F.B.F.), and NSF grants: NSF-IIS-1318386 and NSF-DBI-1458557 (V.B.). The TEM facility is supported in part by NIH award number S10OD023527. Work in the Law laboratory was supported by a Salk Innovation Grant and by the Rita Allen Foundation Scholars Program. H.Y.C. is an Investigator of the Howard Hughes Medical Institute.

Author contributions S.W., K.M.T., V.B., B.R., H.Y.C. and P.S.M. conceived and designed the study. K.M.T. and S.W. performed experiments. N.N. and S.W. performed data analysis. N.N., R.R., J.A.L., B.L., A.A. and M.H. analysed sequencing data. U.R. developed image analysis pipeline. M.R.C., X.C., J.M.G. and H.Y.C. provided support for ATAC-seq and ATAC-seq experiments and analysis. C.C. and J.A.L. performed long-range optical mapping. H.K., R.G.W.V., M.Y., B.R. and V.B. provided analytic support. M.E., J.S., Y.D., W.Z., N.J., J.H., Z.Y., R.H. and F.B.F. provided experimental support. S.W., K.M.T., V.B. and P.S.M. wrote the manuscript with feedback from all authors.

Competing interests P.S.M., H.Y.C. and R.G.W.V. are co-founders of Boundless Bio, Inc. and serve as consultants. V.B. is a co-founder, and has equity interest in Boundless Bio, Inc. and Digital Proteomics, LLC, and receives income from DP. The terms of this arrangement have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies. Boundless Bio, Inc. and Digital Proteomics, LLC were not involved in the research presented here. K.M.T. and N.N. became employees of Boundless Bio, Inc. after the paper was accepted for publication.

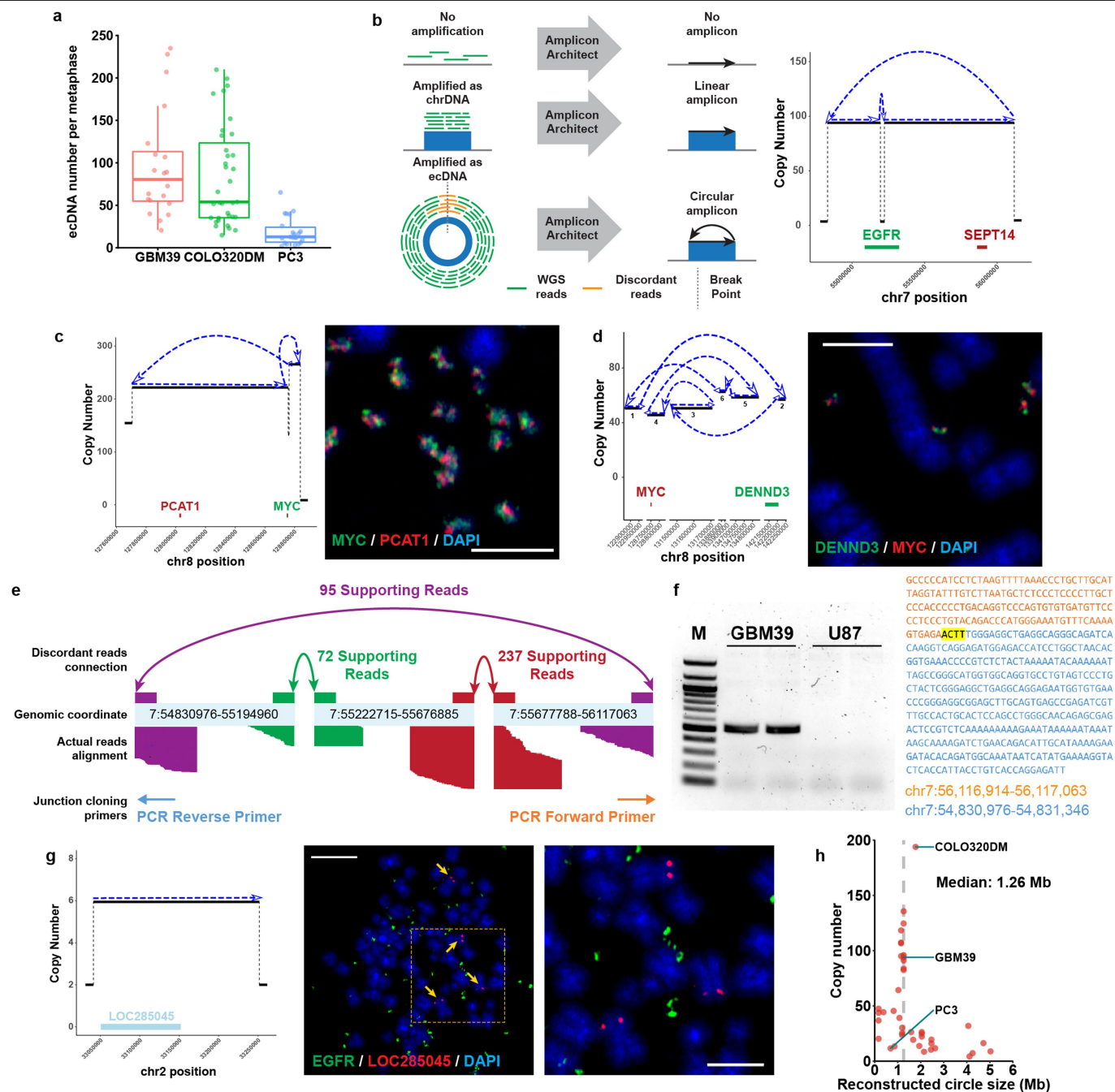
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1763-5>.

Correspondence and requests for materials should be addressed to H.Y.C., B.R., V.B. or P.S.M.

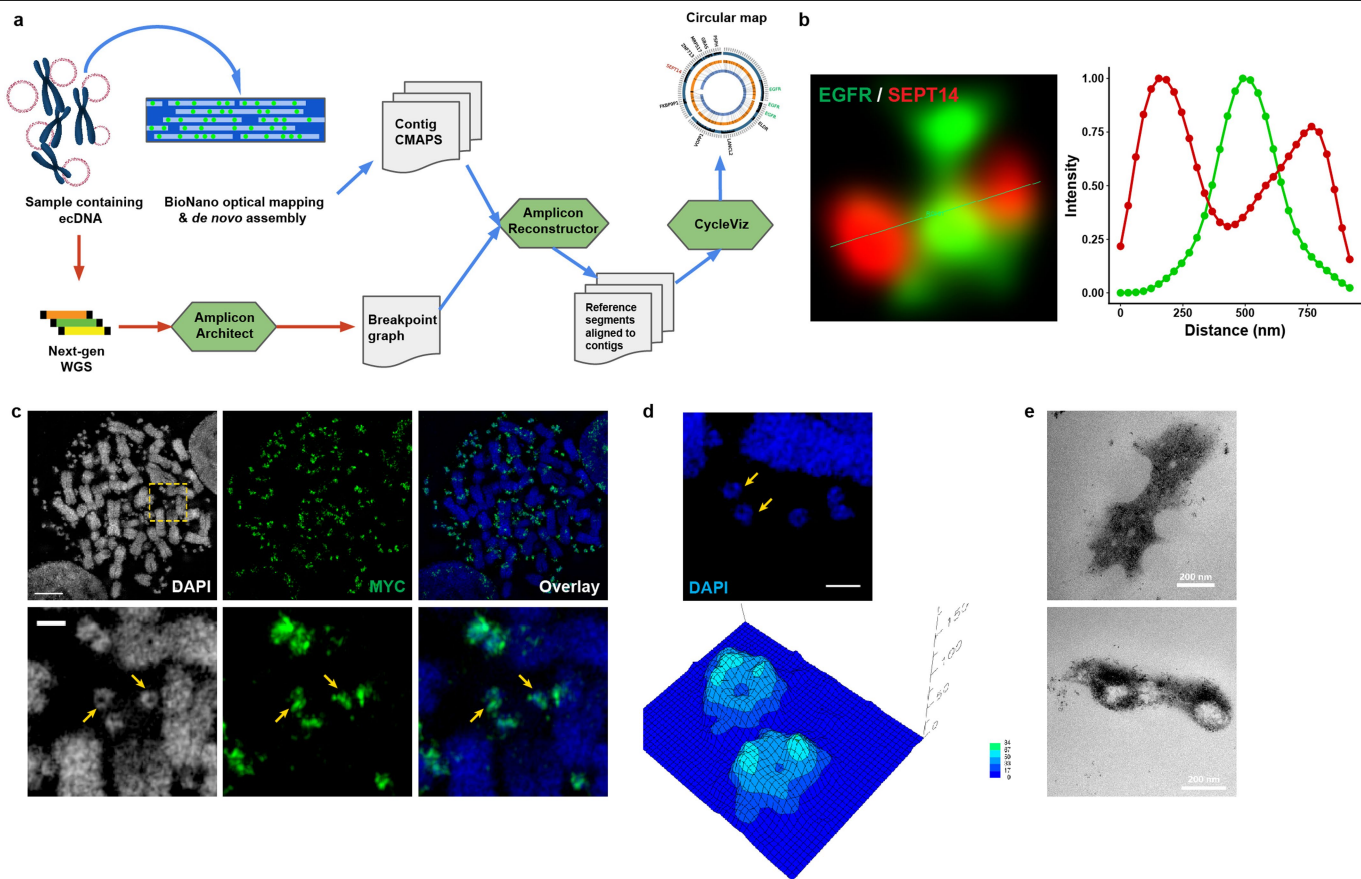
Peer review information *Nature* thanks Tony Papenfuss, Lothar Schermelleh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



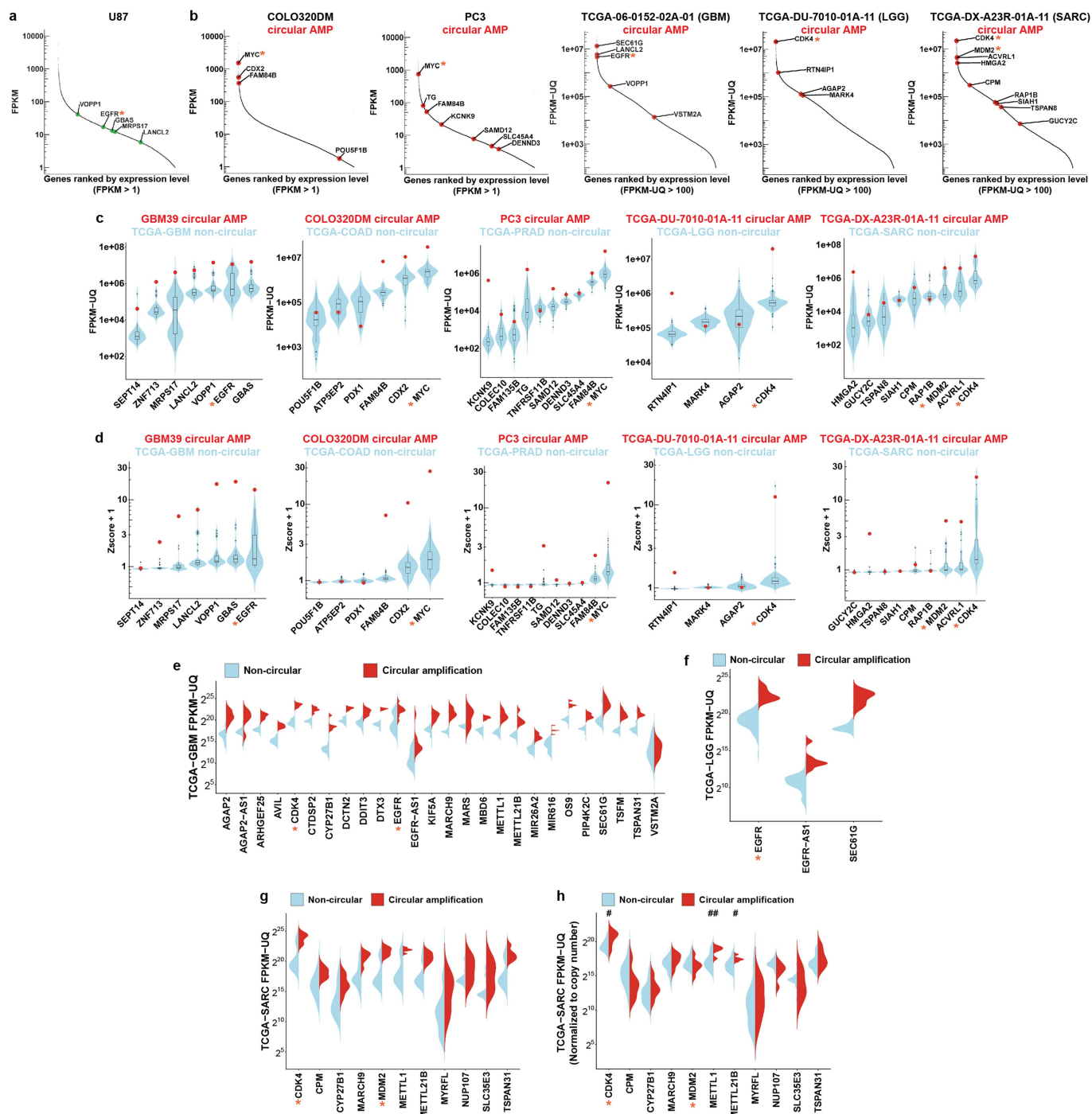
Extended Data Fig. 1 | Characterization of ecDNA structure by WGS. a, ecDNA number per metaphase in GBM39, COLO320DM and PC3 cell lines. Box plots are as in Fig. 2g. At least 20 metaphase spreads from 3 biologically independent samples were counted. **b**, Left, depiction of amplification status classified by AmpliconArchitect. Right, representative AmpliconArchitect of the EGFR circular amplicon in GBM39 cells. Arrows represent the orientation of the assembled contig. **c**, Circular amplicon in COLO320DM cells and double FISH of MYC and PCAT1 validating the amplicon structure. Scale bar, 5 μ m. **d**, Circular amplicon in PC3 cells and double FISH validating the structure and co-existence of DENND3 and MYC in the same ecDNA. Scale bar, 5 μ m. **e**, A detailed AmpliconArchitect-reconstructed schema showing the junctions and hg19 coordinates of ecDNA in GBM39 cells, and the number of paired-end discordant

reads to support the reconstruction. **f**, PCR cloning (left) and Sanger sequencing validation (right) of the ecDNA circular junction in GBM39 cells using the primers in **d**. Exact sequence and BLAT result are shown on the right. The highlighted 4-bp nucleotides were overlaps of the two DNA segments. An ecDNA-free GBM cell line U87 was used as a negative control. M, 100-bp DNA ladder. Data are representative of three independent experiments. See Supplementary Fig. 1 for source data. **g**, Representative linear amplicon breakpoint graph in GBM39 cells (left), with FISH validation of its chromosomal loci (right). Scale bars, 10 μ m (left) and 5 μ m (right). **h**, Size and copy number of 41 reconstructed circular structures in 37 cancer cell lines. All imaging experiments were repeated at least three times, with similar results.



Extended Data Fig. 2 | Characterization of ecDNA structure by optical mapping and imaging. **a**, Pipeline to integrate WGS and BioNano optical mapping. CMAPS denotes a contig mapping and analysis package. **b**, Intensity profile plot of the double FISH of EGFR and SEPT14 in GBM39 cells. **c**, FISH validating MYC-containing ecDNA in COLO320DM cells visualized by 3D-SIM.

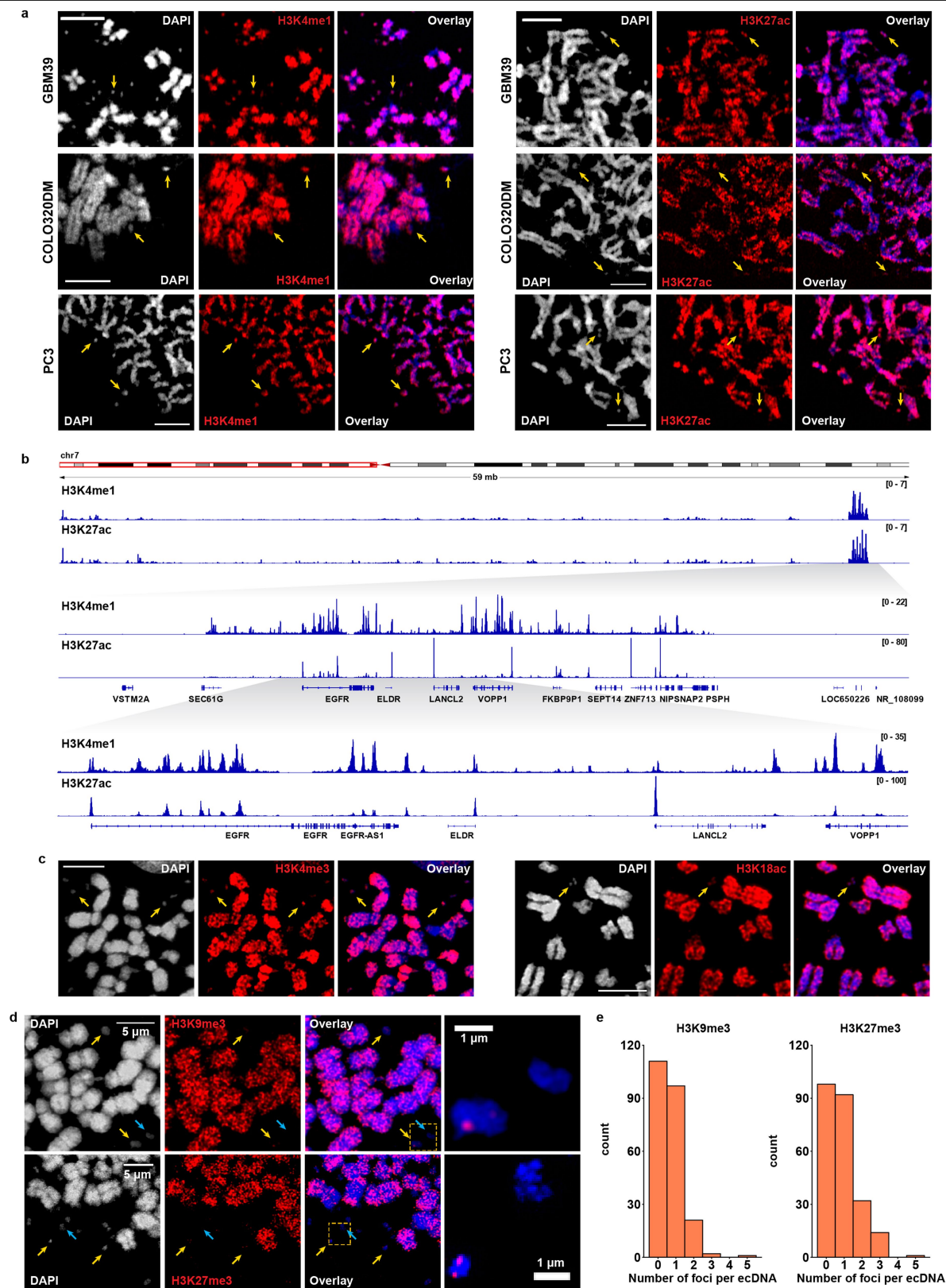
Scale bars, 5 μ m (top) and 1 μ m (bottom). **d**, Three-dimensional reconstruction showing the circular structure of two individual ecDNA structures from 3D-SIM (arrows). The height in the contour map indicates the signal intensity of DAPI. Scale bar, 1 μ m. **e**, TEM of GBM39 ecDNA. Scale bars, 200 nm. All imaging experiments were repeated at least three times, with similar results.



Extended Data Fig. 3 | Genes on ecDNA are highly expressed.

a, Transcriptome in the U87 GBM cell line, which lacks ecDNA. Green data points represent the same genes that are found on ecDNA in the GBM39 cell line. **b**, ecDNA gene expression levels within the transcriptome of COLO320DM and PC3 cells, and selected TCGA samples. Red dots represent genes located on ecDNA (circular amplification genes). **c**, ecDNA gene expression (red data points) in GBM39 cells, COLO320DM cells, PC3 cells, one TCGA-LGG sample (TCGA-DU-7010-01A-11) and one TCGA-SARC sample (TCGA-DX-A23R-01A-11), compared to non-circular genes in the TCGA-GBM ($n = 36$ biologically independent samples), TCGA-COAD ($n = 52$ biologically independent samples), TCGA-PRAD ($n = 120$ biologically independent samples), TCGA-LGG ($n = 96$ biologically independent samples) and TCGA-SARC ($n = 36$ biologically

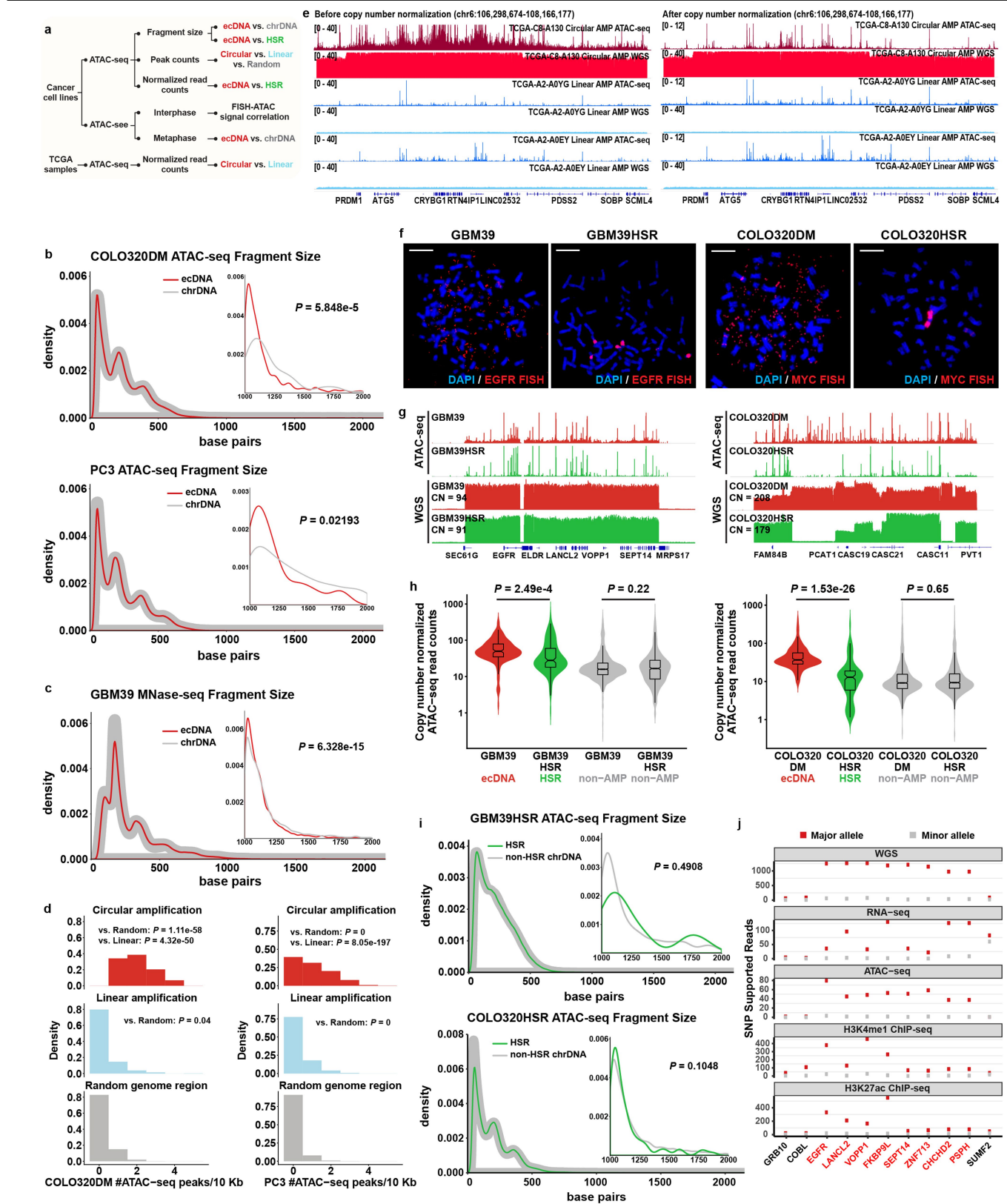
independent samples) cohorts, respectively. **d**, Z-score of the gene expression values in **b**. Z-scores were plotted as +1 to avoid negative values during \log_{10} transformation. For TCGA samples in **b** and **c**, genes on circular amplicons are highlighted as red data points. **e–g**, Expression of circular amplified and non-circular genes in the TCGA-GBM, TCGA-LGG and TCGA-SARC cohorts. **h**, Normalized gene expression by copy number in the TCGA-SARC cohort ($CDK4$, $P < 0.028$; $METTL1$, $P = 0.007$; $METTL21B$). $P = 0.024$, two-sided Wilcoxon rank-sum test. Asterisks indicate key oncogenes. Violin plots show the overall distribution of data points. Box plots are as in Fig. 2g. Every gene in each amplicon type was analysed from at least five biologically independent samples in **e–h**.



Extended Data Fig. 4 | Histone modifications on ecDNA. a,

Immunofluorescence staining of active histone marks H3K4me1 and H3K27ac in metaphase GBM39, COLO320DM and PC3 cells. Scale bars, 5 μ m. **b**, H3K4me1 and H3K27ac ChIP-seq in cycling GBM39 cells. Magnified area demonstrates the ecDNA region. **c**, Immunofluorescence staining of active histone marks H3K4me3 and H3K18ac in metaphase GBM39 cells. Scale bars, 5 μ m.

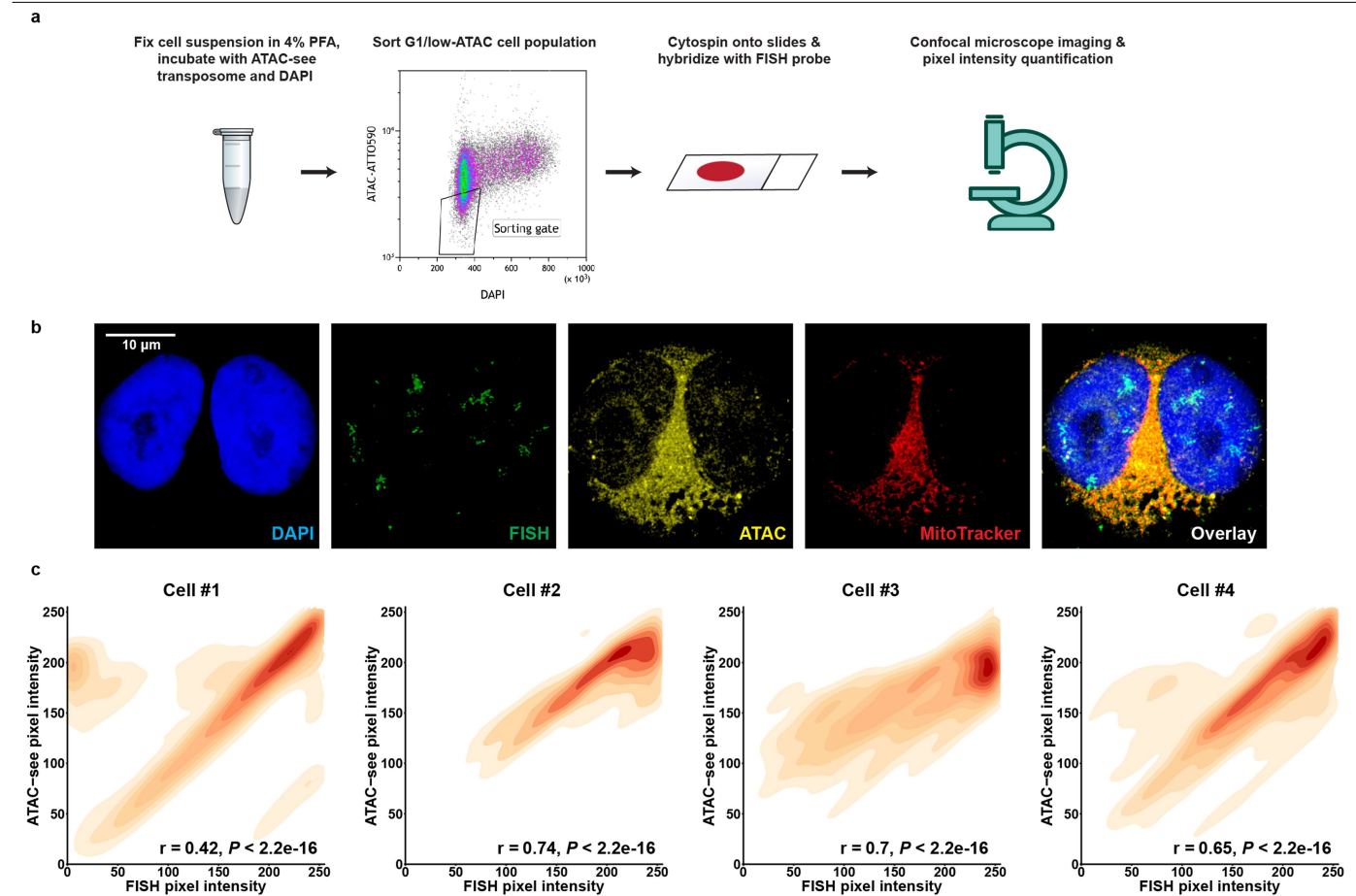
d, Immunofluorescence staining of inactive histone marks H3K9me3 and H3K27me3 in metaphase GBM39 cells. Yellow arrows indicate positive foci, blue arrows indicate ecDNA without foci. **e**, Quantification of H3K9me3 and H3K27me3 foci per ecDNA in GBM39 cells in metaphase. All imaging experiments were repeated at least three times, with similar results.



Extended Data Fig. 5 | See next page for caption.

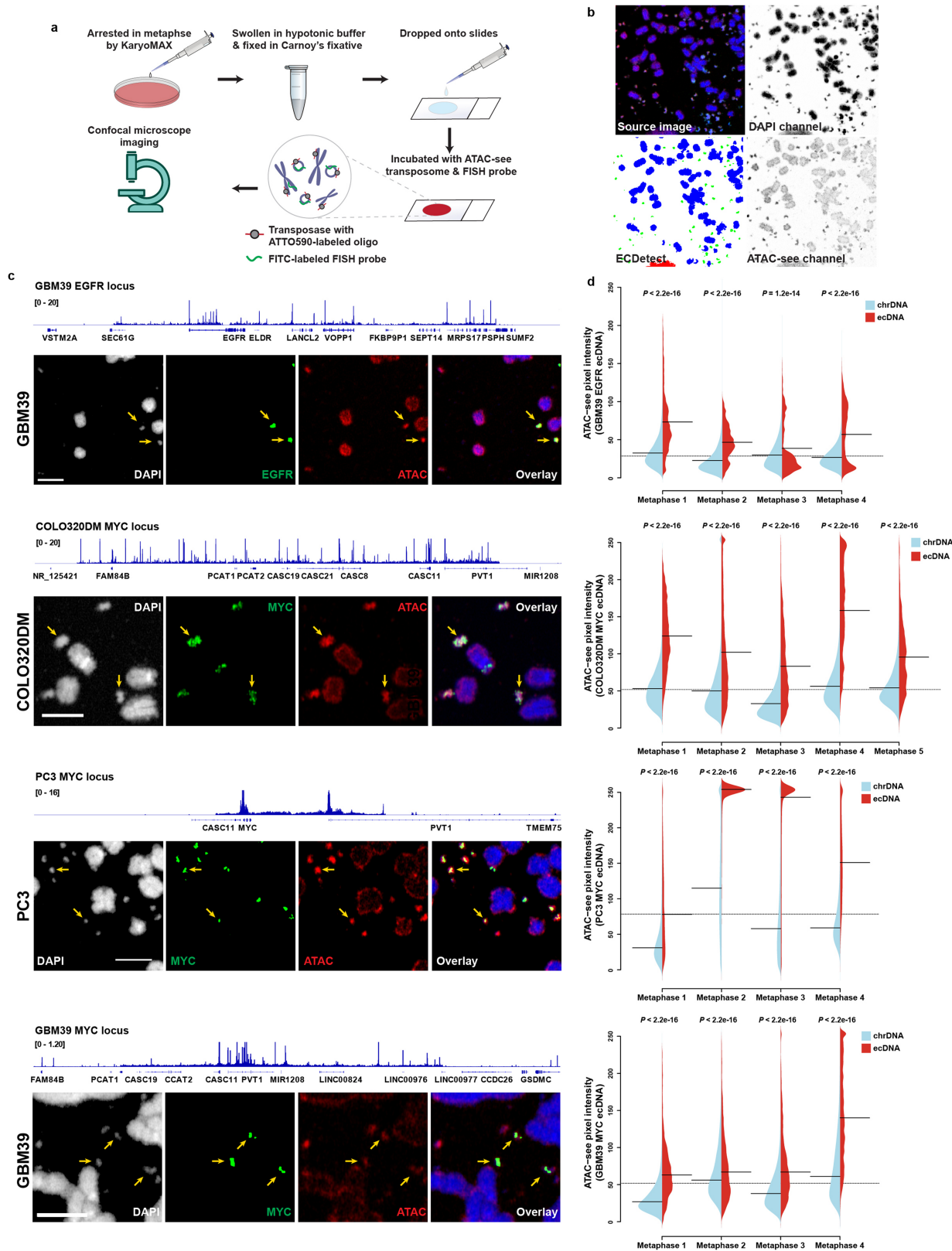
Extended Data Fig. 5 | ecDNA chromatin compaction. **a**, Workflow to characterize the chromatin accessibility of ecDNA. **b**, Global and long (>1 kb) ATAC-seq read length distribution comparing ecDNA and chrDNA in COLO320DM (88 ecDNA and 987 chrDNA long fragments) and PC3 (39 ecDNA and 108 chrDNA long fragments) cells ($n=2$ biologically independent samples, showing one of the representative results). P values determined by two-sided Kolmogorov–Smirnov test. **c**, Distribution of global and long (>1 kb) MNase-seq fragment lengths in GBM39 cells (2,699 ecDNA and 18,942 chrDNA long fragments; $n=2$ biologically independent samples, showing one of the representative results). P value determined by two-sided Kolmogorov–Smirnov test. **d**, ATAC-seq peak number per 10 kb comparing random genome regions (313,762 windows in COLO320DM and PC3 cells), linear amplification (470 windows in COLO320DM, 15,186 windows in PC3 cells), and circular amplification regions (44 windows in COLO320DM, 510 windows in PC3 cells; $n=2$ biologically independent samples). P values determined by Kruskal–Wallis rank-sum test. **e**, ATAC-seq and WGS tracks of TCGA samples comparing circular and linear amplified regions, before (left) and after (right)

normalization to copy number. **f**, Representative FISH from three replicates showing amplicon location in GBM39, GBM39HSR, COLO320DM and COLO320HSR metaphase cells. Scale bars, 10 μm . **g**, ATAC-seq and WGS tracks of the amplified region in GBM39, GBM39HSR, COLO320DM and COLO320HSR cells. CN, copy number. **h**, Normalized ATAC-seq read counts (10-kb bin) by copy number comparing ecDNA and HSR regions (GBM39/HSR amplicon, 134 windows; COLO320DM/HSR amplicon, 157 windows; non-amplicon, 1,000 windows). P values determined by two-sided Dunn’s test. Violin plots show the overall distribution of data points. Box plots are as in Fig. 2g. **i**, Distribution of global and long (>1 kb) ATAC-seq read lengths comparing HSR and non-HSR chrDNA in GBM39HSR (15 ecDNA and 640 chrDNA long fragments) and COLO320HSR (102 ecDNA and 4,554 chrDNA long fragments) cells ($n=2$ biologically independent samples, showing one of the representative results). P value determined by two-sided Kolmogorov–Smirnov test. **j**, Number of single nucleotide polymorphism (SNP) supported reads from the major allele (containing ecDNA) and minor allele in GBM39 cells from multiple sequencing technologies. Circular amplified region (ecDNA) is marked in red.



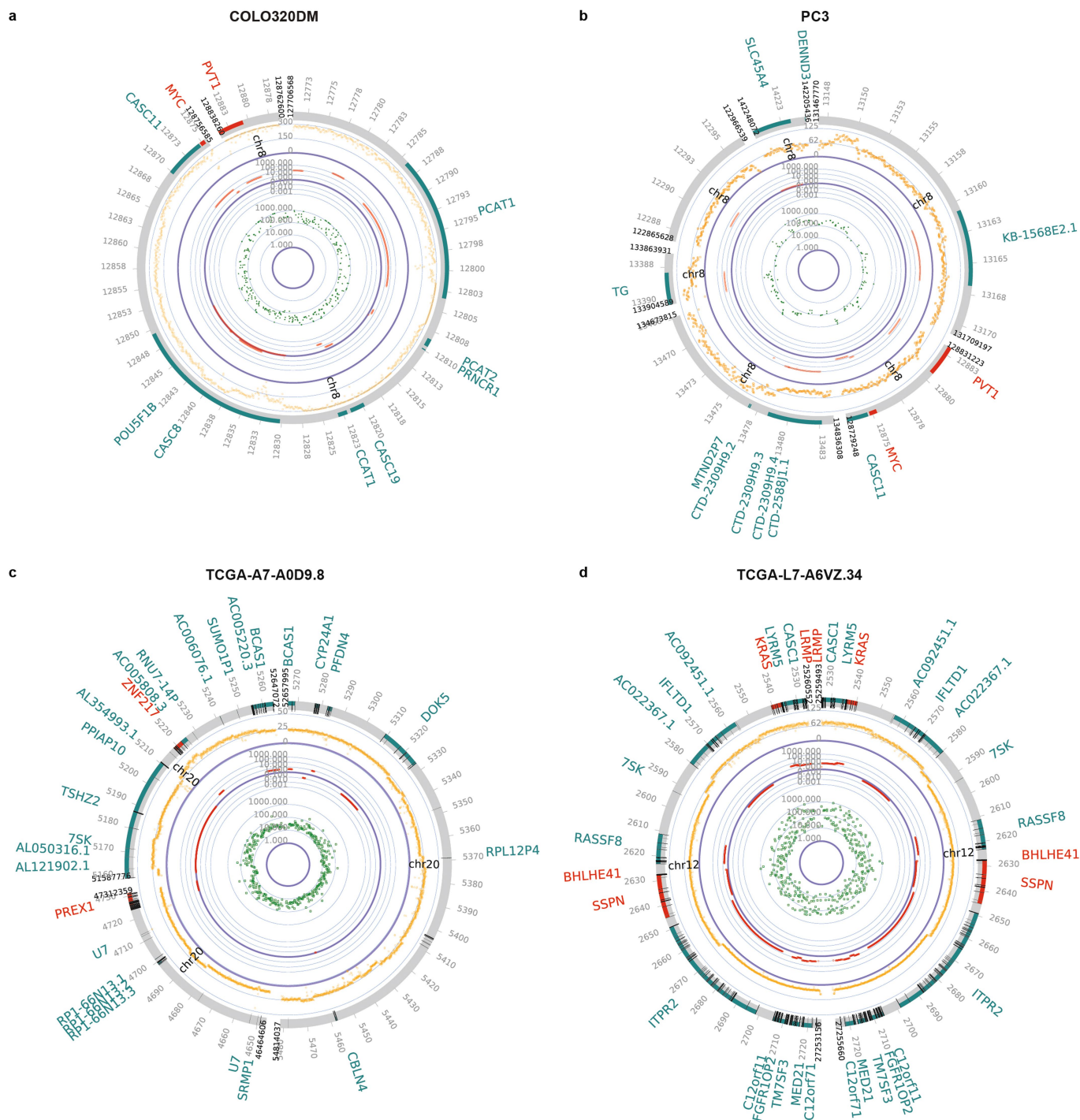
Extended Data Fig. 6 | ecDNA is highly accessible in early interphase chromatin. a, Workflow to evaluate the accessibility of ecDNA in interphase cells. **b,** Representative images of FISH, ATAC-seq and MitoTracker Deep Red

FM signal colocalization in COLO320DM cells. **c,** Pearson correlation of FISH signal pixel intensity and ATAC-seq signal pixel intensity in four representative single cells. At least 27,000 pixels were analysed for each cell.



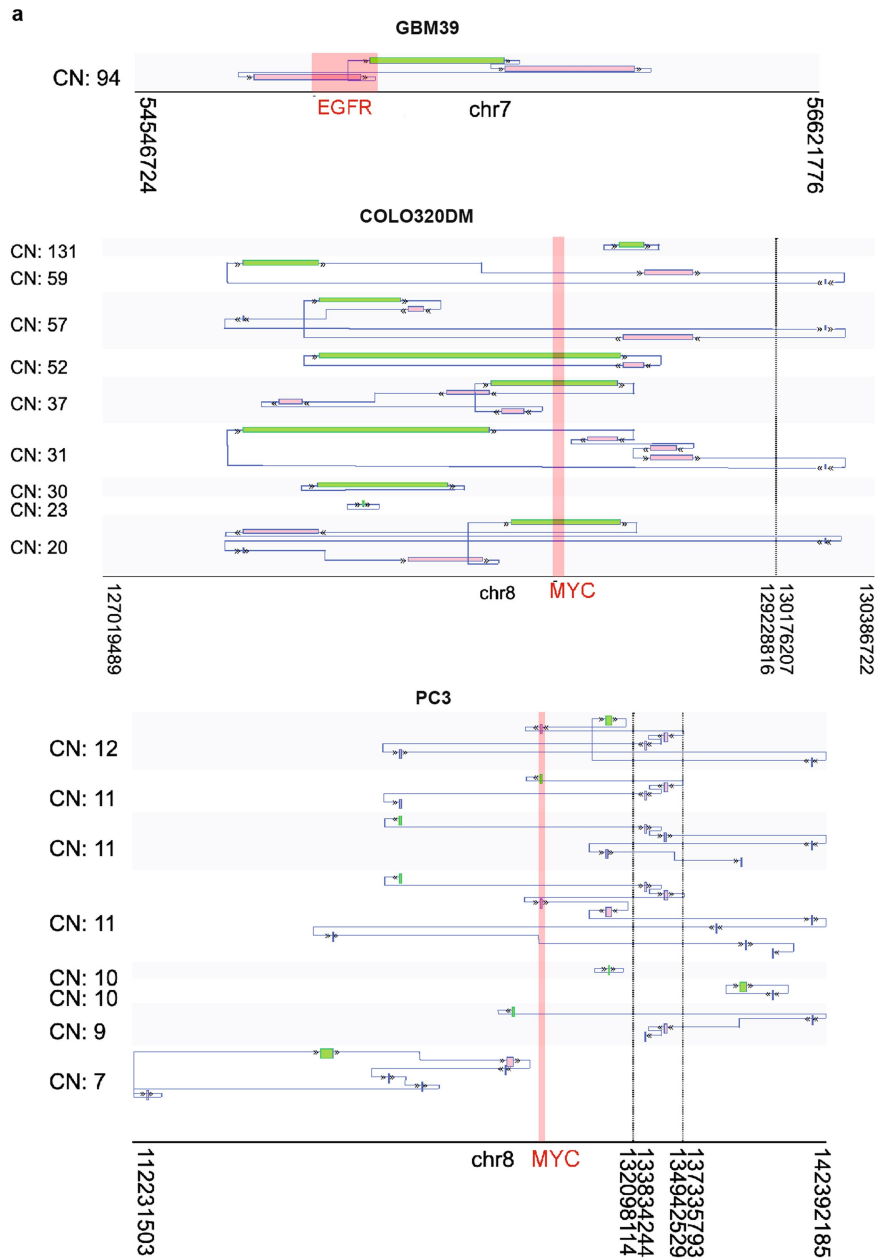
Extended Data Fig. 7 | ATAC-seq visualization of ecDNA accessibility in metaphase chromatin. **a**, The strategy of applying ATAC-seq to DNA in cells in metaphase. **b**, Image analysis pipeline, showing ecDNA and chrDNA segmentation of the DAPI channel. The pixel intensity of ATAC-seq channel was measured. **c**, ATAC-seq tracks and corresponding representative images of FISH and ATAC-seq. Scale bars, 5 μ m. **d**, Quantification of ATAC-seq pixel

intensity of ecDNA versus chrDNA from at least four independent metaphase spreads. Violin plots show the overall distribution of data points. The dashed line across the plot indicates the global mean value. The solid black lines inside each split violin plot indicate the mean of each dataset. P values determined by two-sided Z-test.



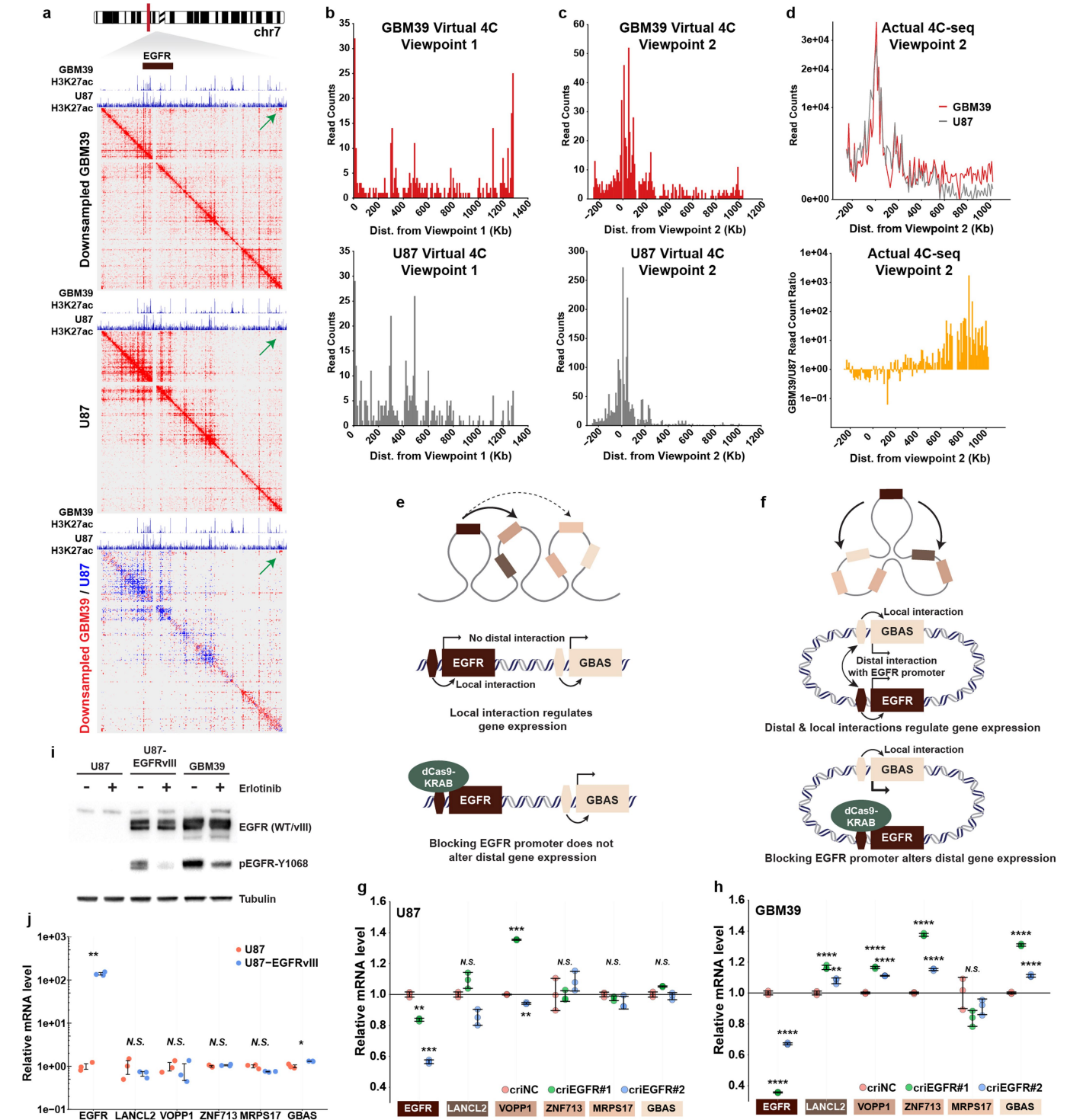
Extended Data Fig. 8 | Circular plots for ecDNA. a–d. Composite circular plots displaying WGS, RNA-seq and ATAC-seq of ecDNA. For COLO320DM and PC3 cells with multiple versions of reconstructed structures, only one

representative structure is shown. For TCGA samples (c, TCGA-A7-A0D9, breast invasive carcinoma; d, TCGA-L7-A6VZ, oesophageal carcinoma), the ATAC-seq data point represents the highest signal within a 1-kb window.



Extended Data Fig. 9 | Reconstructed ecDNA structures. a, Examples of selected potential amplicons reconstructed from AmpliconArchitect in GBM39, COLO320DM and PC3 cells. For each potential amplicon, the average copy number of the segments is listed. The starting segment of the structure is

outlined in green. From the starting segment, the structure can be traced by following the arrows to find the next genomic segment of the structure. Some structures have a circular path (that is, can return to the starting segment by following the arrows), which represents potential ecDNA structure.



Extended Data Fig. 10 | Circularization of ecDNA enables novel DNA interaction. **a**, Chromatin interaction heat maps comparing GBM39 with U87 cells, generated from PLAC-seq/HiChIP analyses using H3K27ac as the anchor. The GBM39 ecDNA region was downsampled to a comparable level of U87 to normalize for copy number. Contrast heat map shows the differential interaction. Green arrows indicate the increased corner reads in the GBM39 ecDNA junctional region but not in the U87 chrDNA locus, demonstrating ecDNA circularity. **b, c**, Virtual 4C read counts from viewpoints 1 (ecDNA junction) and 2 (*EGFR* promoter), respectively. **d**, Actual 4C-seq read counts, and the read count ratio of GBM39 to U87 from viewpoint 2. **e, f**, Models depicting local and distal interactions with the *EGFR* promoter and proposed model for CRISPR interference masking of the *EGFR* promoter. **g, h**, qPCR

analysis of gene expression in regions proximal and distal to *EGFR*. Data are mean \pm s.e.m.; $n = 3$; each data point represents three technical replicates from one representative result. *criEGFR*, CRISPR interference of *EGFR*; *criNC*, CRISPR interference negative control. $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$, one-way ANOVA. N.S., not significant. **i**, Exogenous expression of *EGFR* variant III in U87 cells (U87-*EGFR*vIII) and the activation of *EGFR* signalling was confirmed by western blot. Experiment was repeated three times, with similar results. See Supplementary Fig. 1 for source data. **j**, qPCR analysis of *EGFR*-neighbouring gene expression in U87 cells, with and without ectopic overexpression of *EGFR* variant III. Data are mean \pm s.e.m.; $n = 3$; each data point represents three technical replicates from one representative result. *GBAS*, $*P = 0.038$; *EGFR*, $**P = 0.003$; Welch's *t*-test.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

AmpliconReconstructor was used to integrate the optical mapping contigs with AA based WGS-reconstructions (<https://github.com/jluebeck/AmpliconReconstructor>). CycleViz (<https://github.com/jluebeck/CycleViz>) was utilized to generate Fig. 1c. Circular plots in Fig. 4a and Extended Data Fig. 8 were generated by CircularPlot (<https://github.com/namphuon/CircularPlot>). WGS, RNA-seq, and ATAC-seq data were obtained from the TCGA project (<http://cancergenome.nih.gov>). Confocal images were captured and analyzed by Zeiss ZEN Black (version 2.3 SP1 FP3), ZEN blue (version 2.5) and Leica Lightning Imaging Information Extraction Software (3.5.5.19976). Structured Illumination reconstructions were performed using Softworx version 6.5.2. For BioNano data analysis, raw images were processed, and long DNA molecules were detected and digitized by BioNano image-processing and analysis software AutoDetect31. Optical maps were generated by transforming the raw images into raw BNX files using the IrysView software system. The BNX files output from the BioNano instrument were then assembled into optical map contigs using the BioNano Irys assembly pipeline (v5122, default parameters)

Data analysis

Data were analyzed by R program. ATAC-seq data on metaphase chromosome was analyzed by ECDetect. ATAC-seq data on interphase nuclei was analyzed by ImageJ. 3D rendering of SIM image was performed by Softworx software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data are available in SRA database and released on 9/13/2019, with BioProject #PRJNA506071.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For imaging, qPCR and Western blot, experiments were performed at least 3 times in biologically independent replicates. We included all available TCGA samples for analysis, and excluded those with less than 5 samples.
Data exclusions	No data we generated were excluded. For public TCGA data, data with N < 5 were excluded.
Replication	All experiments were repeated at least 3 times with similar results to ensure reproducibility.
Randomization	The present study uses cultured human cancer cells for experimental intervention, and randomization is not applicable in this case.
Blinding	All experimental data were acquired by instrumental analysis objectively without anthropic bias, and blinding is not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

ChIP-seq:
 anti-H3K27ac: Active Motif, Cat# 39685, Lot# 31718018, 5 µg per ChIP
 anti-H3K4me1: Active Motif, Cat# 39297, Lot# 01518002, 10 µl per ChIP
 anti-CTCF: Abcam, Cat# ab70303, Lot# GR3218438-7, 5 µg per ChIP
 anti-SMC3: Abcam, Cat# ab9263, Lot# GR3221084-5, 5 µg per ChIP

PLAC-seq:
 anti-H3K27ac: Diagenode, Cat# C15200184-50, Lot# 001-13, 2.5 µg per ChIP

Western blot:
 anti-EGFR: EMD Millipore, Cat# 06-847, Lot# 3016636, dilution 1:5000
 anti-phospho-EGFR: CST, Cat# 3777S, Lot# 13, dilution 1:1000
 anti-Tubulin: CST, Cat# 2125S, Lot# 9, dilution 1:2000
 secondary anti-rabbit IgG antibody: CST, Cat#7074S, Lot# 26, dilution 1:2000

Immunofluorescence:

anti-H3K4me1: CST, Cat# 5326, Lot# 1, dilution 1:800
 anti-H3K27ac: CST, Cat# 8173, Lot# 1, dilution 1:300
 anti-H3K4me3: Diagenode, Cat# C15410003, Lot# A1052D, dilution 1:200
 anti-H3K18ac: Diagenode, Cat# C15410139, Lot# A1460D, dilution 1:200
 anti-H3K9me3: Active Motif, Cat# 39765, Lot# 16513004, dilution 1:400
 anti-H3K27me3: Active Motif, Cat# 39155, Lot# 31218021, dilution 1:500

Validation

All antibodies are validated to react with corresponding human antigens. Citation data are acquired from CiteAb database.

ChIP-seq:

anti-H3K27ac: Active Motif, Cat# 39685, 16 citations
 anti-H3K4me1: Active Motif, Cat# 39297, 35 citations
 anti-CTCF: Abcam, Cat# ab70303, 42 citations
 anti-SMC3: Abcam, Cat# ab9263, 45 citations

PLAC-seq:

anti-H3K27ac: Diagenode, Cat# C15200184-50, 2 citations

Western blot:

anti-EGFR: EMD Millipore, Cat# 06-847, 62 citations
 anti-phospho-EGFR: CST, Cat# 3777S, 328 citations
 anti-Tubulin: CST, Cat# 2125S, 288 citations
 secondary anti-rabbit IgG antibody: CST, Cat#7074S, 3952 citations

Immunofluorescence:

anti-H3K4me1: CST, Cat# 5326, 21 citations
 anti-H3K27ac: CST, Cat# 8173, 56 citations
 anti-H3K4me3: Diagenode, Cat# C15410003, 89 citations
 anti-H3K18ac: Diagenode, Cat# C15410139, 1 citations
 anti-H3K9me3: Active Motif, Cat# 39765, 17 citations
 anti-H3K27me3: Active Motif, Cat# 39155, 168 citations

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

COLO320DM, COLO320HSR and PC3 cells were purchased from ATCC. GBM39 and GBM39HSR are patient-derived cell lines that have been previously described in Turner et al. Nature. 2017.

Authentication

Cell lines were obtained from ATCC and therefore were not authenticated. GBM39 and GBM39HSR cells were authenticated to the same as in Turner et al. Nature. 2017.

Mycoplasma contamination

All cell-lines were tested negative for mycoplasma

Commonly misidentified lines
(See [ICLAC](#) register)

None of the cell line is listed in ICLAC Register of Misidentified Cell Lines

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA506071> (Note: We have uploaded all the FASTQ files to SRA. However, SRA does not provide temporary access link)

Files in database submission

GBM39KT_SMC3_rep2_r1.fq.gz, GBM39KT_SMC3_rep2_r2.fq.gz, GBM39KT_SMC3_rep1_r1.fq.gz,
 GBM39KT_SMC3_rep1_r2.fq.gz, GBM39KT_Input_rep2_r1.fq.gz, GBM39KT_Input_rep2_r2.fq.gz,
 GBM39KT_Input_rep1_r2.fq.gz, GBM39KT_H3K27ac_rep2_r2.fq.gz, GBM39KT_H3K4me1_rep2_r1.fq.gz,
 GBM39KT_H3K27ac_rep1_r1.fq.gz, GBM39KT_H3K27ac_rep2_r1.fq.gz, GBM39KT_CTCF_rep1_r1.fq.gz,
 GBM39KT_Input_rep1_r1.fq.gz, GBM39KT_H3K4me1_rep1_r1.fq.gz, GBM39KT_CTCF_rep1_r2.fq.gz,
 GBM39KT_H3K4me1_rep1_r2.fq.gz, GBM39KT_CTCF_rep2_r2.fq.gz, GBM39KT_CTCF_rep2_r1.fq.gz,
 GBM39KT_H3K4me1_rep2_r2.fq.gz, GBM39KT_H3K27ac_rep1_r2.fq.gz

Genome browser session
(e.g. [UCSC](#))

We provide the IGV session file for reviewers. (igv_session_for_reviewer.xml)

Methodology

Replicates	Two biological replicates and perform library preparation and sequencing at the same time.
Sequencing depth	CTCF (2 replicates): Total mapped reads 4.0 & 5.2M, Total unique reads 3.9 & 5.0M H3K27ac (2 replicates): Total mapped reads 9.7 & 8.8M, Total unique reads 9.6 & 8.7M H3K4me1 (2 replicates): Total mapped reads 11.3 & 8.5M, Total unique reads 11.2M & 8.3M SMC3 (2 replicates): Total mapped reads 9.3 & 11.2M, Total unique reads 9.1 & 10.9M Input (2 replicates): Total mapped reads 7.3 & 11.5M, Total unique reads 7.1 & 11.4M
Antibodies	See Method section in the manuscript
Peak calling parameters	macs2 callpeak -t treatment.bam -c input.bam -f BAM -n sample --outdir sample_macs2 -g hs -p 1e-2 --nomodel --shift 0 --extsize 100 --keep-dup all -B --SPMR
Data quality	We used the default setting of MACS2, i.e. the minimum fold change range we used from MACS2 is the default [5-50].
Software	Bowtie2 for mapping, MarkDuplicates from Picard tools v1.119 to remove PCR duplicates, Samtools v1.3.1, MACS2 2.1.2 for peak calling.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	We prepared single cell suspensions of COLO320DM cells, which were fixed in 1% formaldehyde solution in PBS for 10 minutes. Cells were centrifuged and resuspended in lysis buffer (as previously described). Five million cells were stained with 100 nM Atto-594-labeled Tn5 transposase in TD Buffer (as previously described) for 30 minutes at 37C. Cells were then washed in excess of PBS and centrifuged, followed by DAPI staining in PBS for 30 minutes. Cells were washed with an excess of PBS before being resuspended in 1X PBS for FACS sorting. Detailed method was described in Chen et al., Nature Methods 2016.
Instrument	SONY SH800 cell sorter
Software	SONY SH800 software was used for cell sorting. The image in the schema in Extended Data Fig. 5a was generated by Beckman Coulter Kaluza software
Cell population abundance	About 20% cells in the population were sorted. The cell sorting in the current study was not for purification, but enrichment for low ATAC-seq signal population. Detailed method was described in Chen et al., Nature Methods 2016.
Gating strategy	We used FSC-A/SSC-A to locate the major cell population, and FSC-H/FSC-W, DAPI-A/DAPI-H to gate single cell. The population with relatively low ATAC-seq signal was gated and sorted. Detailed method was described in Chen et al., Nature Methods 2016.
<input checked="" type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM

<https://doi.org/10.1038/s41586-019-1768-0>

Thomas C. R. Miller¹, Julia Locke¹, Julia F. Greiwe¹, John F. X. Diffley² & Alessandro Costa^{1*}

Received: 16 May 2019

Accepted: 27 September 2019

Published online: 20 November 2019

In preparation for bidirectional DNA replication, the origin recognition complex (ORC) loads two hexameric MCM helicases to form a head-to-head double hexamer around DNA^{1,2}. The mechanism of MCM double-hexamer formation is debated. Single-molecule experiments have suggested a sequential mechanism, in which the ORC-dependent loading of the first hexamer drives the recruitment of the second hexamer³. By contrast, biochemical data have shown that two rings are loaded independently via the same ORC-mediated mechanism, at two inverted DNA sites^{4,5}. Here we visualize MCM loading using time-resolved electron microscopy, and identify intermediates in the formation of the double hexamer. We confirm that both hexamers are recruited via the same interaction that occurs between ORC and the C-terminal domains of the MCM helicases. Moreover, we identify the mechanism of coupled MCM loading. The loading of the first MCM hexamer around DNA creates a distinct interaction site, which promotes the engagement of ORC at the N-terminal homodimerization interface of MCM. In this configuration, ORC is poised to direct the recruitment of the second hexamer in an inverted orientation, which is suitable for the formation of the double hexamer. Our results therefore reconcile the two apparently contrasting models derived from single-molecule experiments and biochemical data.

Genome replication in eukaryotes is tightly controlled to ensure that chromosomes are copied only once per cell cycle. Before cells enter S phase, two copies of an MCM heterohexameric helicase are loaded onto duplex DNA to form a double hexamer^{1,2,6,7}. Double hexamers mark origin DNA that can support replisome assembly and—once activated—the MCMs unwind DNA, providing the template for replicative polymerases to perform bidirectional replication. Inhibiting the formation of the double hexamer is a major pathway for preventing re-replication and maintaining genome stability^{8,9}. MCM loading requires (i) DNA association of the ORC and Cdc6, (ii) opening of a DNA gate in MCM (which is stabilized by Cdt1^{1,2,10,11}) and (iii) ATP hydrolysis by MCM to close the gate around DNA^{12,13}.

A model for the loading of a first MCM helicase onto DNA has previously been proposed on the basis of biochemical^{4,14}, single-molecule¹⁵ and cryo-electron microscopy (cryo-EM) analysis^{10,11,16,17,18}. ORC first encircles and bends origin DNA¹⁸. Upon Cdc6—and then MCM—recruitment, ORC threads the double helix through the DNA gate of MCM (between Mcm2 and Mcm5), which leads to the formation of an ORC–Cdc6–Cdt1–MCM intermediate (hereafter, OCCM) that encircles DNA^{17–19}. In this complex, the C-terminal face of an ORC–Cdc6 ring engages the C-terminal face of a notched MCM ring. Notably, OCCM has previously been observed in the presence of a slowly hydrolysable ATP analogue that allows MCM recruitment to DNA but not the hydrolysis-dependent MCM ring closure that is required to complete loading^{4,14,17}.

The mechanism for the recruitment of a second MCM ring and the formation of the double hexamer is debated. According to a previous single-molecule study³, the first MCM ring is loaded by one ORC

complex and then drives the recruitment of the second ring. However, biochemical evidence⁴ has shown that the same elements in MCM are required for recruiting both the first and second rings. Combined with the observation that two inverted ORC binding sites promote efficient MCM loading⁵, these data support a model in which two distinct engagement events between ORC and DNA symmetrically load two MCM rings via the same mechanism. Explaining the formation of the head-to-head double hexamer is critical for understanding how the symmetry of bidirectional replication is established.

Entire origins of replication visualized

To elucidate the steps that lead to the formation of the double hexamer, we took an electron microscopy approach to visualize the entire ATPase-dependent MCM loading reaction, reconstituted *in vitro* with purified yeast proteins. As a substrate for helicase loading, we used linear DNA that contains the *ARS1* origin sequence. This sequence features a high-affinity ORC binding site (the autonomous replicating sequence (ARS) consensus sequence, ACS), which maps 42 bp upstream of an inverted low-affinity site (B2)^{5,20} (Fig. 1a). Because double hexamers passively slide on duplex DNA^{1,6,7,21}, we were unable to retain MCM particles on *ARS1* DNA tethered to streptavidin-coated magnetic beads. We reasoned that nucleosomes, which naturally flank *ARS1* *in vivo*²², might function as roadblocks to limit the sliding of double hexamers (Fig. 1b). Indeed, nucleosome-capped *ARS1*—but not naked *ARS1*—retains loaded MCM after a high-salt wash. The same result is obtained by chromatinizing naturally occurring *ARS1* flanking sequences (Extended Data Fig. 1a)

¹Macromolecular Machines Laboratory, Francis Crick Institute, London, UK. ²Chromosome Replication Laboratory, Francis Crick Institute, London, UK. *e-mail: alessandro.costa@crick.ac.uk

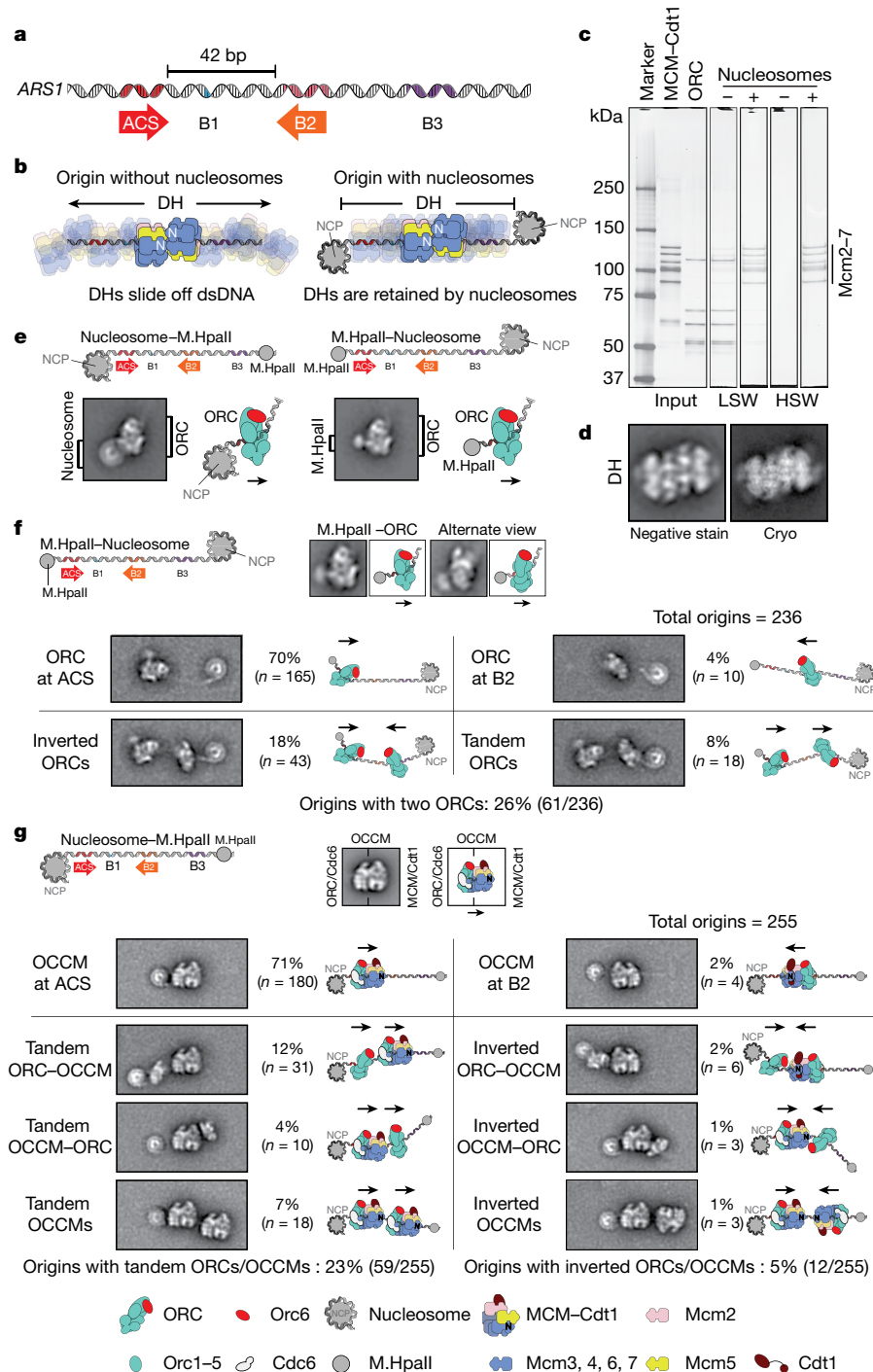


Fig. 1 | In silico reconstitution of MCM recruitment to origins. **a**, Linear structure of the *ARS1* origin of replication in yeast. ACS and B2 are inverted high- and low-affinity binding sites for ORC, respectively. **b**, Schematic of double-hexamer (DH) capture on DNA using nucleosomes as roadblocks. dsDNA, double-stranded DNA. NCP, nucleosome core particle. **c**, Nucleosomes work as roadblocks that prevent the linear diffusion of double hexamers. Nucleosome-decorated, but not naked *ARS1* DNA substrates retain MCM particles after both low- and high-salt washes (LSW and HSW, respectively) in a bead-based DNA pulldown assay. For gel source data, see Supplementary Fig. 1. **d**, Loaded MCM helicases form double hexamers. **e**, ORC complexes interact

with the *ARS1* ACS. Two-dimensional class averages of ORC bound to asymmetric origins show the ORC average is in close proximity to the ACS-proximal roadblock. In **e–g**, black arrows indicate the orientation of the C-terminal MCM-recruitment interface of ORC. **f**, In silico reconstitution of ORC binding to asymmetric origins shows multiple modes of ORC interaction. Representative reconstituted origins are shown; schematics indicate the directionality of ORC binding. The frequency of each class of ORC binding event is indicated as a percentage of total origins analysed ($n = 236$). **g**, In silico reconstitution as in **f**, showing MCM recruitment to origins in ATP γ S ($n = 255$). Mcm3, 4, 6, 7 denotes Mcm3, Mcm4, Mcm6 and Mcm7.

or strong-positioning Widom sequences that are biochemically more tractable than the native DNA (Fig. 1c).

Bead-free helicase-loading reactions analysed by electron microscopy show efficient formation of the double hexamer, as well as class

averages that contain isolated ORC, MCM-Cdt1, nucleosome or nucleosome close to ORC (Fig. 1d, Extended Data Fig. 1b). To establish whether the efficient formation of the double hexamer requires a specific interaction between the ORC and the nucleosome, we repeated the

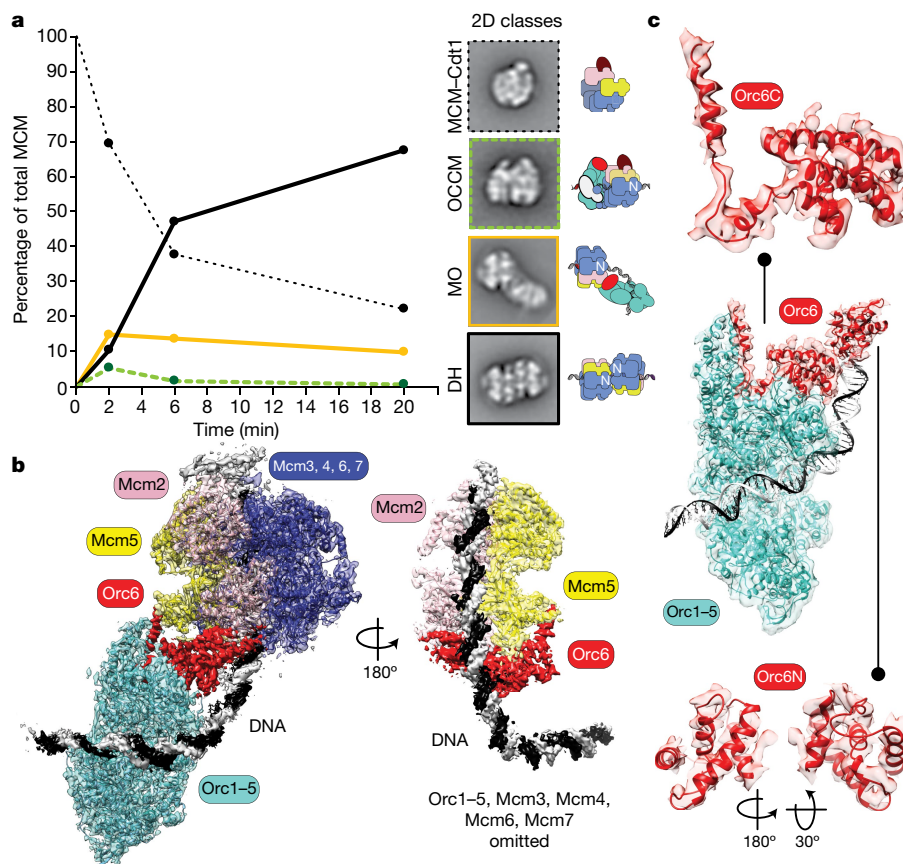


Fig. 2 | Time-resolved helicase-loading experiments lead to the identification of MCM loading intermediates. **a**, A helicase-loading time course analysed by negative-stain electron microscopy shows that the proportion of MCM molecules in the loading-competent MCM–Cdt1 species decreases as the loaded double-hexamers count increases. At 2 min, the OCCM loading intermediate peaks. At the same time, a second loading intermediate—which evidently contains MCM and ORC (MO)—can be identified. **b**, Cryo-EM structure and atomic model of the MO intermediate. The model was obtained

by real-space-refining docked coordinates of MCM (PDB 6EYC (ref. ²⁴)), ORC (PDB 5ZR1 (ref. ¹⁸)) and an N-terminal Orc6 homology model. Orc6 (red) bridges between Orc1–5 and MCM, contacts the N-terminal face of MCM and latches across the A domains of Mcm2 and Mcm5. MCM and ORC are DNA-bound, and span more than half of the *ARS1* origin sequence; the DNA is bent and solvent-exposed between the ORC and MCM complexes. **c**, Views of N-terminal and C-terminal Orc6 in the context of the ORC in the MO complex.

MCM loading reaction while replacing nucleosome caps with M.HpaII methyltransferase adducts (Extended Data Fig. 1c). We find that MCM loading efficiency is unperturbed, which indicates that the nucleosomes in our assay primarily serve as roadblocks that prevent the sliding of the double hexamer. To discriminate between ORC binding at ACS versus B2, we designed asymmetric substrates that contain either a nucleosome upstream of ACS and an M.HpaII 75 bp downstream of the B2 element (nucleosome–M.HpaII), or the inverse (M.HpaII–nucleosome) (Extended Data Fig. 1d). Electron microscopy imaging of ORC binding revealed virtually identical ORC views, which mapped in very close proximity to the nucleosome or the M.HpaII adduct—depending on which neighbored the ACS (Fig. 1e). This result demonstrates that the preference of ORC for the high-affinity ACS site is unaffected by nucleosomes.

As duplex DNA is flexible, single-particle 2D averaging did not capture the whole context of ORC bound to the *ARS1* origin; precluding, for example, the identification of multiple ORC binding events on a single origin. To understand origin architecture during the formation of the double hexamer, we established an *in silico* reconstitution approach (which we term ‘ReconSil’) to generate signal-enhanced views of entire replication origins (Extended Data Fig. 2). To this end, we computationally re-assembled origins by substituting raw particles with class averages, re-oriented using particle coordinates from the original micrographs and alignment parameters from 2D classification. The criteria we used to ensure that particles mapped

on the same origin DNA are detailed in the Supplementary Methods and Extended Data Fig. 2. To validate this approach, we measured a set of 226 M.HpaII–nucleosome ORC-bound origins reconstituted *in silico*, and found an average length of 141 bp (s.d. of 11 bp). The expected distance between the terminal M.HpaII and nucleosome roadblocks is 143 bp.

Our ReconSil experiments show 70% (165 out of 236) of origins were bound by a single ORC at the ACS (Fig. 1f), 4% (10 out of 236) were bound only at the B2 site and 26% (61 out of 236) were simultaneously engaged by two ORCs. The distinctive 2D views of ORC enable the assignment of the relative ORC orientations (Fig. 1f). In 18% of cases (43 out of 236; only 70% of the 61 doubly populated origins), two ORCs are oriented with their MCM-interacting domains facing one another, as predicted by a symmetric mechanism for the formation of the double hexamer^{5,23} (Fig. 1f).

Next, we performed a helicase recruitment assay using ATPγS to capture OCCM complexes on asymmetric nucleosome–M.HpaII origins. Similar to origins bound by ORC alone, 71% (180 out of 255) of origins reconstituted *in silico* are engaged by a single OCCM at the ACS, and only 2% (4 out of 255) by a single OCCM at B2 (Fig. 1g, Extended Data Fig. 2d). The remaining 71 origins were simultaneously bound either by 2 OCCM complexes or by OCCM and ORC, although they rarely (12 out of 255; 17% of the 71 doubly populated origins) had ORC and OCCM in the inverted orientation that is predicted for a simple symmetric mechanism of loading (Fig. 1g).

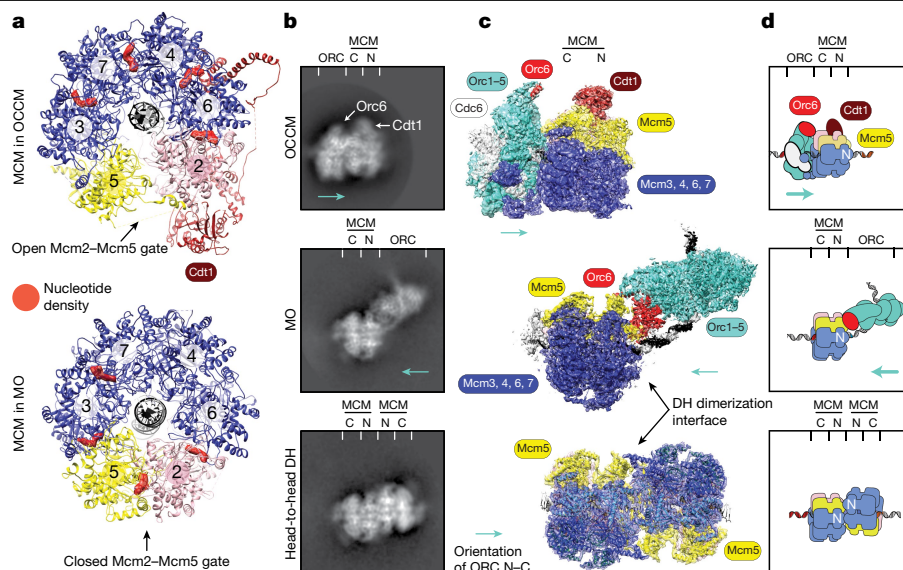


Fig. 3 | The MO intermediate contains a post-catalytic closed MCM ring.

a, MCM in OCCM contains a notched ring with an opening that spans nearly the entirety of Mcm2 and Mcm5¹⁷. ATPase interfaces at Mcm3-Mcm7, Mcm7-Mcm4, Mcm4-Mcm6 and Mcm6-Mcm2 are ATPγS-bound in the OCCM structure. MCM in MO contains a closed Mcm2-Mcm5 gate. As observed in the post-catalytic DNA-loaded double hexamer⁷, nucleotide density (probably of ADP) can be found at the Mcm6-Mcm2, Mcm2-Mcm5, Mcm5-Mcm3 and Mcm3-Mcm7

ATPase interfaces. **b–d**, Cryo-EM 2D classes (**b**), 3D structures (**c**) and schematics (**d**) comparing the OCCM, MO and double-hexamer assemblies. In OCCM (top panels), the C-terminal domains of ORC and MCM (C) form the interaction interface and both Cdc6 and Cdt1 are present. In MO (middle panels), ORC binds to the N-terminal (N) dimerization interface of MCM through Orc6. Cdc6 and Cdt1 are absent. The double hexamer (bottom panels) forms a head-to-head, symmetrical dimer through interactions between its N-terminal domains.

Time-resolved imaging of MCM loading

To investigate the mechanism through which the double hexamer is formed, we performed an MCM-loading time-course assay in ATP for electron microscopy imaging at 2, 6 and 20 min after MCM addition. As expected, by quantifying the percentage of particles that contribute to MCM-containing classes, we observed a decrease in MCM-Cdt1 complexes (to 69%, 38% and 22% of total MCM-containing classes at 2, 6 and 20 min, respectively). This change was compensated for by an increase in double hexamers (10%, 47% and 67% at 2, 6 and 20 min, respectively). A subset of OCCM particles (5%) appeared at 2 min, and nearly vanished after 20 min (1%) (Fig. 2a). Thus, OCCM forms in ATP and disappears as double hexamers assemble, which provides strong support for its role as a bona fide loading intermediate.

A fourth species peaked at 2 min (15% of MCM-containing particles) (Fig. 2a) and gradually tailed off. This species apparently contains one ORC complex, poised about 90° offset from the central channel of a single MCM ring (Fig. 2a). The composition and architecture of this species suggested that it may be a loading intermediate, formed downstream of OCCM and perhaps involved in the recruitment of the second MCM hexamer. We reasoned that the continued presence of this species throughout the time course may have been the result of MCM depletion preventing the formation of the double hexamer. Accordingly, by adjusting MCM concentration and loading time, we were able to enrich for this MCM-ORC species (Extended Data Fig. 3a).

To understand the nature of this MCM-ORC interaction, we solved the 4.4 Å resolution cryo-EM structure (Fig. 2b, Extended Data Figs. 3b–g, 4, Extended Data Table 1). Rigid-body-fitting of available ORC DNA¹⁸ and MCM DNA^{6,7,24} structures confirms that this DNA-bound complex contains an MCM hexamer that interacts with ORC (Fig. 2b).

In our structure, MCM and ORC are connected by two elements. The first is duplex DNA that becomes bent as it runs through the ORC cavity; this DNA remains solvent-exposed as it occupies a gap in the protein structure and traverses the entire MCM channel (Fig. 2b). A cut-through view reveals DNA spanning 88 bp (Fig. 2b). A second element that tethers ORC to MCM involves the C-terminal domain of Orc6

and a neighbouring protein module that was previously unresolved. We postulated that this element could be the N-terminal domain of Orc6, which has so far eluded structural characterization. Consistent with published bioinformatics analysis¹⁹, our homology modelling of N-terminal yeast Orc6 reveals a TFIIb-like fold (Extended Data Fig. 5). Docking of the N-terminal Orc6 model into the cryo-EM map resulted in an excellent fit (Fig. 2c). Combining this information with the ORC and MCM structures, we refined a near-complete atomic model of the full DNA-bound complex (Supplementary Video 1), which we refer to as MCM-ORC (MO).

Two aspects of the structure deserve attention. First, MO lacks Cdc6 and Cdt1 (Fig. 3, Extended Data Fig. 6). ATP hydrolysis is known to promote Cdt1 ejection and closure of the MCM ring^{4,11,13,15}. Accordingly, we find that the gate between Mcm2 and Mcm5 (Mcm2-Mcm5) in MO is closed, and that the nucleotide occupancy of MCM matches that of a loaded (post-catalytic) double hexamer, with nucleotide-binding sites at the Mcm4-Mcm6 and Mcm7-Mcm4 interfaces empty, and the Mcm6-Mcm2, Mcm2-Mcm5, Mcm5-Mcm3 and Mcm3-Mcm7 interfaces bound by nucleotide (probably ADP)⁷ (Fig. 3a). Moreover, molecular docking reveals that MCM in MO is virtually identical to an MCM hexamer in the double hexamer^{6,7} (Supplementary Video 2). Thus, MO contains a single-loaded helicase ring, which supports our hypothesis that MO occurs after the OCCM intermediate¹⁵.

Second, ORC in MO binds DNA in an inverted configuration with respect to the upstream ACS element and interacts with the N-terminal (rather than the C-terminal) side of the MCM ring, in contrast to ORC in OCCM¹⁷ (Fig. 3b–d). This mode of binding requires closure of the MCM gate, so that the N- and C-terminal elements of Orc6 can latch across Mcm2 and Mcm5 (the subunits of the DNA gate) (Fig. 2b, Extended Data Fig. 5d, e). In doing so, Orc6 also engages a protein surface that is involved in MCM homodimerization in a double hexamer¹⁶ (Fig. 3).

N-terminal Orc6 promotes MCM loading

To investigate whether the interaction between MCM and ORC that is observed in MO has a role in the formation of the double hexamer,

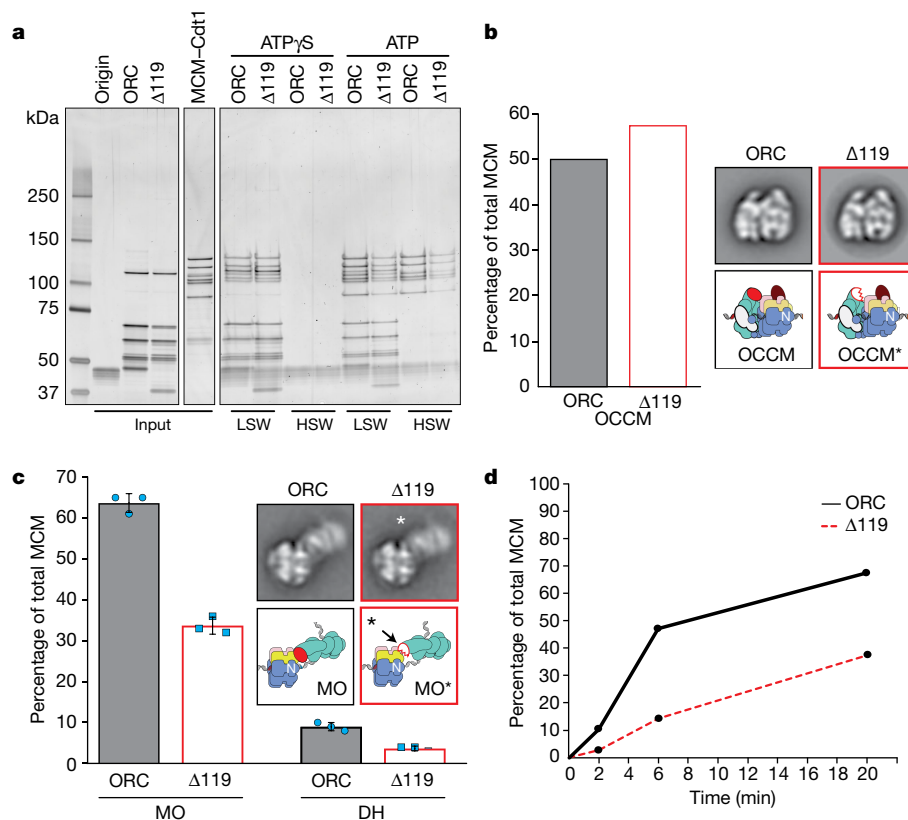


Fig. 4 | N-terminal Orc6 truncation reduces MCM loading but not recruitment. **a**, A bead-based DNA pulldown assay demonstrates that Δ 119 recruits MCM in ATP γ S as efficiently as wild-type ORC; however, Δ 119 is less efficient at MCM loading, as observed both in low and high-salt wash conditions. For gel source data, see Supplementary Fig. 1. **b**, According to negative-stain electron microscopy experiments, OCCM formation is as efficient for Δ 119 as it is for wild-type ORC. **c**, The efficiency of MO formation

drops by about 50% when using Δ 119 instead of wild-type ORC, and the efficiency of double-hexamer formation decreases by more than half. Bar chart shows mean \pm s.d., $n = 3$ independent experiments. *Negative-stain 2D classes of MO containing wild-type ORC and Δ 119 show a subtle change in ORC orientation; this is shown most clearly in Supplementary Video 3. **d**, In an electron-microscopy-based time-course experiment, the formation of double hexamers markedly slows when using Δ 119 instead of wild-type ORC.

we sought to impair the interface between Orc6 and Mcm2–Mcm5. The ORC complex lacking Orc6 (Orc1–5) can recruit MCM to origins in ATP γ S, but releases MCM from DNA in ATP γ S. We therefore performed a subtler Orc6 truncation that lacks the N-terminal domain (which precludes Mcm2 engagement in MO) but retains the Cdt1 interaction that is essential for MCM loading^{9,25,26}. To this end, we generated an ORC variant that lacks the N-terminal 119 amino acids of Orc6 (hereafter, Δ 119). Using a bead-based pulldown assay, we showed that Δ 119 recruits MCM in ATP γ S to wild-type levels (Fig. 4a) and efficiently forms OCCM, as shown by electron microscopy analysis (Fig. 4b). However, the ATP-promoted formation of MO decreased by about 50% with Δ 119 compared to wild-type ORC (Fig. 4c). Notably, Δ 119 MOs have an altered ORC conformation (Supplementary Video 3). Thus, removing just one component at the interface of the MCM–ORC interaction negatively affects MO formation.

Consistent with the notion that MO is a bona fide loading intermediate, when ORC is replaced by Δ 119 (i) the retention of double hexamers on DNA beads drops in an *in vitro* helicase-loading reaction subjected to low or high-salt washes (Fig. 4a); (ii) double hexamer counts decrease by 2.6-fold in the MO-formation electron microscopy assay (Fig. 4c); and (iii) the formation of double hexamers is substantially slowed in a negative-stain electron-microscopy time-course experiment (Fig. 4d). The impaired ability of Δ 119 to load double hexamers may explain a previous observation that an Orc6 that lacks amino acids 1–73 cannot complement an Orc6 deletion *in vivo*⁹.

MCM double hexamer formation is coupled

We used time-resolved electron microscopy to show that OCCM is a bona fide reaction intermediate that forms at ACS sites, on the path to loading the first MCM hexamer. To further understand the recruitment of the first hexamer, we performed an analogous cryo-EM experiment using an early time point and conditions optimized to capture a pre-OCCM state. Aside from reproducing the expected ATP–OCCM in vitreous ice, we also identified a 2D class referred to as OC–MC, in which ORC–Cdc6 contacts, but is yet to fully engage MCM–Cdt1 to form OCCM (Fig. 5a, Extended Data Fig. 7a, b). In this class, duplex DNA is poised for threading into the MCM channel, as it is bent by ORC and aligned with the Mcm2–Mcm5 gate. We confirmed this observation by comparing origins reconstituted *in silico* that contained OC–MC and OCCM particles, which were obtained in our time-course experiment for the formation of the double hexamer (Extended Data Fig. 7c, d). This finding reveals an initial helicase-engagement mode on the path to the first MCM loading.

As ORC in MO recognizes a closed MCM ring at the MCM dimerization interface, we postulated that this state might reflect a downstream ORC binding event, on the path to the recruitment of the second ring. Indeed, superposition of the ORC complexes in the MO and OCCM structures reveals that ORC-induced DNA-bending in MO creates sufficient space for the inverted ORC to accept a second MCM ring via the same interactions that form OCCM⁴ (Extended Data Fig. 8a, b). In this model, the Mcm2–Mcm5 gate in the second incoming MCM would align to the solvent-exposed bent DNA in MO, and therefore be perfectly positioned for DNA threading into the MCM channel.

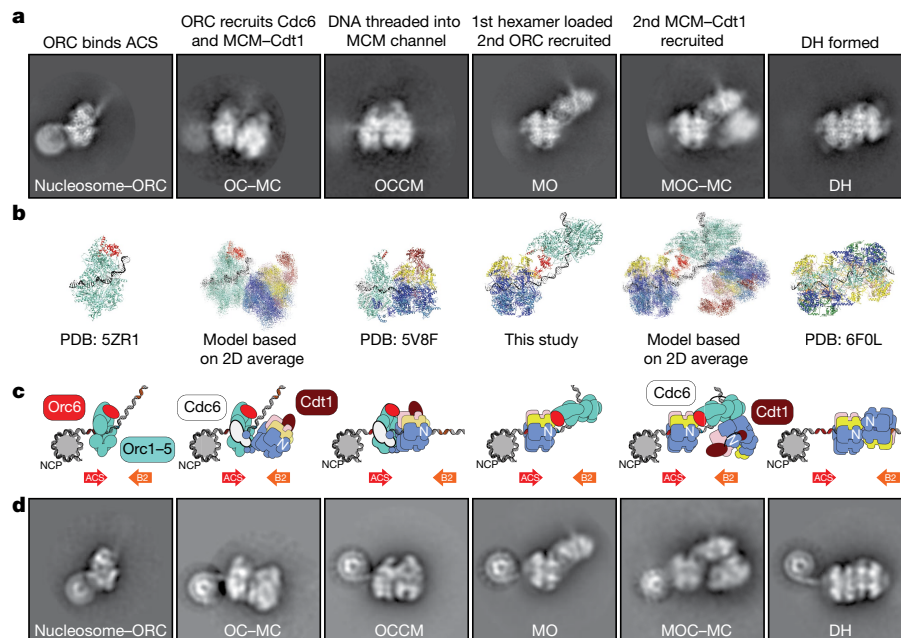


Fig. 5 | The MCM double-hexamer formation reaction visualized by electron microscopy. a, Cryo-EM 2D averages reveal the steps that lead to the formation of the double hexamer. From left to right: ORC binds at the ACS (next to a nucleosome) and bends the DNA. ORC recruits the first MCM ring. DNA is threaded through the Mcm2–Mcm5 gate, forming the OCCM. ATP hydrolysis leads to Cdc6 and Cdt1 release, and Mcm2–Mcm5 gate closure, which promotes a second ORC binding event at the N-terminal face of MCM, forming the MO intermediate. In this configuration, ORC engages a second MCM via an

OC–MC-like intermediate (MOC–MC); here, the second MCM is aligned for symmetrical helicase loading. Finally, the double hexamer is formed around DNA. **b, c**, The same reaction represented with atomic models or noisy projections for models generated from 2D averages (**b**) or with cartoons (**c**). **d**, First and second MCM recruitment and loading on nucleosome–M.HpaII origins, visualized by negative-stain electron microscopy and ReconSil. These images show the position of each of the licensing intermediates, relative to the ACS-flanking nucleosome.

In negative-stain electron microscopy experiments, we successfully captured this predicted state by supplementing MO with MCM–Cdt1 to visualize the engagement of the second MCM hexamer (Extended Data Fig. 8c). In a parallel effort, subclassification of MO images in the cryo-EM dataset enabled us to identify the same molecular species showing second-ring recruitment (Fig. 5a, Extended Data Fig. 8d). The clear DNA visualization that was possible with cryo-EM revealed that this species indeed contains DNA that is bent and aligned with the Mcm2–Mcm5 gate, virtually identical to that in the OC–MC (Fig. 5a, Extended Data Fig. 8d, e, Supplementary Video 4).

Single-molecule studies have shown that the loading of the first and second MCM hexamers require distinct Cdc6 molecules¹⁵. To investigate whether MO-dependent formation of the double hexamer requires Cdc6, we formed MO and immunodepleted Cdc6, before supplementing MOs with MCM–Cdt1. Notably, MOs could not load double hexamers in the absence of additional Cdc6 (Extended Data Fig. 8f). Therefore, we confirm that the loading of the first and second MCM rings occur via the same Cdc6-dependent OC–MC mechanism.

In summary, time-resolved electron microscopy analysis of MCM loading using wild-type proteins and ATP have enabled us to identify intermediates on the path to the formation of the double hexamer. First, ORC, and then Cdc6, bind to high-affinity ACS sites and recruit a first MCM via an interaction between the respective C-terminal domains of ORC and MCM. During this process, the OCCM is formed when bent DNA is threaded into the MCM channel. After the release of Cdc6 and Cdt1, and closure of the MCM ring, an inverted ORC engages the MCM N-terminal homodimerization interface, which leads to the formation of MO. Engagement of a second ORC complex can occur before the release of the first ORC is complete, as shown by an intermediate that contains a loaded MCM flanked by two ORCs (Extended Data Fig. 9). Once engaged, ORC in the MO recruits a second Cdc6 and MCM–Cdt1, which forms an intermediate that comprises a loaded MCM ring and

OC–MC (MOC–MC). ORC therefore bridges a loaded MCM helicase and a second, recruited MCM–Cdt1. As with loading of the first ring, ORC is poised for threading the bent DNA through the Mcm2–Mcm5 gate, which eventually results in the formation of the double hexamer^{1,2}. Our 2D averages can be used as frames to generate a movie of the step-by-step formation of the double hexamer, imaged as it occurs in vitro. Similarly, a molecular morph movie that was generated by interpolating between atomic models of distinct loading intermediates enables visualisation of helicase loading in three dimensions (Fig. 5a–c, Supplementary Video 4).

Recruitment of a second ORC to origin DNA has previously been shown to require a secondary inverted ORC binding site; however, the precise sequence and distance of this binding site from ACS varies between origins^{5,20,27,28}. Therefore, the formation of the double hexamer might require sliding of MCM-loading intermediates^{5,27}.

Using ReconSil, we show that—similar to double hexamers—single MCM helicases can slide on origin DNA^{1,6,7}. In fact, MCM helicases from both MO and MOC–MC intermediates are often found adjacent to the nucleosome, occupying the ACS site that was originally engaged by the first ORC (Fig. 5d, Extended Data Fig. 9). In the MO intermediate, 66 bp of DNA span the DNA entry point into MCM and the ORC-engagement site. Sixty-six base pairs is also the distance that separates the nucleosome positioning sequence from B2 in our construct. Therefore, loaded MCM sliding enables the engagement of inverted ORC at B2, which would otherwise be sterically occluded by the first loaded hexamer on an *ARS1* origin. In this context, ORC could simultaneously engage MCM and B2, increasing its specificity and affinity for the B2 site. This could explain why ORC engagement downstream of the ACS can occur in either orientation when imaged in ORC-binding or MCM-recruitment ATPγS assays (Fig. 1f, g), and why inverted ORC occupancy is assured in the presence of a single loaded MCM. Thus, the formation of MO ensures that the second ORC association occurs with the correct

orientation for the formation of a head-to-head double hexamer. This model has important implications for understanding helicase-loading mechanisms in eukaryotes, in which ORC recognition of origins does not depend on specific DNA sequences²³ (discussed in Supplementary Information).

Our data explain how double-hexamer loading onto replication origins is symmetrical (two inverted ORC assemblies recruit two MCM rings via the same OCCM mechanism^{4,5}) as well as sequential and coordinated (the loading of the first MCM ring drives the loading of the second ring³). Therefore, our model reconciles two apparently contrasting studies on the mechanism of formation of the double hexamer.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1768-0>.

1. Remus, D. et al. Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**, 719–730 (2009).
2. Evrin, C. et al. A double-hexameric MCM2–7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc. Natl Acad. Sci. USA* **106**, 20240–20245 (2009).
3. Tica, S., Friedman, L. J., Ivica, N. A., Gelles, J. & Bell, S. P. Single-molecule studies of origin licensing reveal mechanisms ensuring bidirectional helicase loading. *Cell* **161**, 513–525 (2015).
4. Frigola, J., Remus, D., Mehanna, A. & Diffley, J. F. ATPase-dependent quality control of DNA replication origin licensing. *Nature* **495**, 339–343 (2013).
5. Coster, G. & Diffley, J. F. X. Bidirectional eukaryotic DNA replication is established by quasi-symmetrical helicase loading. *Science* **357**, 314–318 (2017).
6. Abid Ali, F. et al. Cryo-EM structure of a licensed DNA replication origin. *Nat. Commun.* **8**, 2241 (2017).
7. Noguchi, Y. et al. Cryo-EM structure of MCM2–7 double hexamer on DNA suggests a lagging-strand DNA extrusion model. *Proc. Natl Acad. Sci. USA* **114**, E9529–E9538 (2017).
8. Nguyen, V. Q., Co, C. & Li, J. J. Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature* **411**, 1068–1073 (2001).
9. Chen, S. & Bell, S. P. CDK prevents MCM2–7 helicase loading by inhibiting Cdt1 interaction with Orc6. *Genes Dev.* **25**, 363–372 (2011).
10. Zhai, Y. et al. Open-ringed structure of the Cdt1–Mcm2–7 complex as a precursor of the MCM double hexamer. *Nat. Struct. Mol. Biol.* **24**, 300–308 (2017).
11. Frigola, J. et al. Cdt1 stabilizes an open MCM ring for helicase loading. *Nat. Commun.* **8**, 15720 (2017).

12. Coster, G., Frigola, J., Beuron, F., Morris, E. P. & Diffley, J. F. Origin licensing requires ATP binding and hydrolysis by the MCM replicative helicase. *Mol. Cell* **55**, 666–677 (2014).
13. Kang, S., Warner, M. D. & Bell, S. P. Multiple functions for Mcm2–7 ATPase motifs during replication initiation. *Mol. Cell* **55**, 655–665 (2014).
14. Fernández-Cid, A. et al. An ORC/Cdc6/MCM2–7 complex is formed in a multistep reaction to serve as a platform for MCM double-hexamer assembly. *Mol. Cell* **50**, 577–588 (2013).
15. Tica, S. et al. Mechanism and timing of MCM2–7 ring closure during DNA replication origin licensing. *Nat. Struct. Mol. Biol.* **24**, 309–315 (2017).
16. Li, N. et al. Structure of the eukaryotic MCM complex at 3.8 Å. *Nature* **524**, 186–191 (2015).
17. Yuan, Z. et al. Structural basis of MCM2–7 replicative helicase loading by ORC–Cdc6 and Cdt1. *Nat. Struct. Mol. Biol.* **24**, 316–324 (2017).
18. Li, N. et al. Structure of the origin recognition complex bound to DNA replication origin. *Nature* **559**, 217–222 (2018).
19. Bleichert, F., Leitner, A., Aebersold, R., Botchan, M. R. & Berger, J. M. Conformational control and DNA-binding mechanism of the metazoan origin recognition complex. *Proc. Natl Acad. Sci. USA* **115**, E5906–E5915 (2018).
20. Palzkill, T. G. & Newton, C. S. A yeast replication origin consists of multiple copies of a small conserved sequence. *Cell* **53**, 441–450 (1988).
21. Gros, J. et al. Post-licensing specification of eukaryotic replication origins by facilitated MCM2–7 sliding along DNA. *Mol. Cell* **60**, 797–807 (2015).
22. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**, 748–753 (2010).
23. Robinson, N. P. & Bell, S. D. Origins of DNA replication in the three domains of life. *FEBS J.* **272**, 3757–3766 (2005).
24. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D* **74**, 519–530 (2018).
25. Chen, S., de Vries, M. A. & Bell, S. P. Orc6 is required for dynamic recruitment of Cdt1 during repeated MCM2–7 loading. *Genes Dev.* **21**, 2897–2907 (2007).
26. Asano, T., Makise, M., Takehara, M. & Mizushima, T. Interaction between ORC and Cdt1p of *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **7**, 1256–1262 (2007).
27. Warner, M. D., Azmi, I. F., Kang, S., Zhao, Y. & Bell, S. P. Replication origin-flanking roadblocks reveal origin-licensing dynamics and altered sequence dependence. *J. Biol. Chem.* **292**, 21417–21430 (2017).
28. Wilmes, G. M. & Bell, S. P. The B2 element of the *Saccharomyces cerevisiae* ARS1 origin of replication requires specific sequences to facilitate pre-RC formation. *Proc. Natl Acad. Sci. USA* **99**, 101–106 (2002).
29. Nikolov, D. B. et al. Crystal structure of a TFIIB–TBP–TATA-element ternary complex. *Nature* **377**, 119–128 (1995).
30. Balasov, M., Huijbregts, R. P. H. & Chesnokov, I. Role of the Orc6 protein in origin recognition complex-dependent DNA binding and replication in *Drosophila melanogaster*. *Mol. Cell. Biol.* **27**, 3143–3153 (2007).
31. Bleichert, F., Botchan, M. R. & Berger, J. M. Crystal structure of the eukaryotic origin recognition complex. *Nature* **519**, 321–326 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Cryo-EM map and atomic model coordinates for the MO complex have been deposited in the Electron Microscopy Data Bank and Protein Data Bank (PDB), respectively, under the accession codes 6RQC and EMD-4980.

Acknowledgements We thank past and present members of the Costa laboratory for useful discussions; G. Coster, A. McClure, C. Kurat, A. Early and L. Drury for sharing reagents and purification protocols; S. Webb and N. Turner for help with biochemical and electron microscopy experiments; A. Nans for support on the Titan Krios; R. Carzaniga (Electron Microscopy STP) for support on the Tecnai G2 Spirit electron microscope; P. Rosenthal for advice; A. Purkiss and P. Walker (Structural Biology STP) for computational support; and N. Patel, A. Alidoust and D. Patel (Fermentation STP) for yeast cultures. This work was funded jointly

by the Wellcome Trust, MRC and CRUK at the Francis Crick Institute (FC001065 and FC001066). A.C. receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 820102). This work was also funded by a Wellcome Trust Senior Investigator Award (106252/Z/14/Z) and a European Research Council Advanced Grant (669424-CHROMOREP) to J.F.X.D.

Author contributions T.C.R.M. and A.C. conceived the study. T.C.R.M. designed biochemistry experiments. T.C.R.M., J.F.G. and J.L. prepared biochemical reagents and developed the assays. T.C.R.M., J.F.G. and J.L. performed negative-stain imaging. T.C.R.M. and J.L. performed cryo-EM imaging. T.C.R.M. performed all image processing and atomic model building and developed the ReconSil method. J.F.X.D. provided reagents. A.C. supervised the study. T.C.R.M. and A.C. wrote the manuscript with input from J.F.X.D. and the other authors.

Competing interests The authors declare no competing interests.

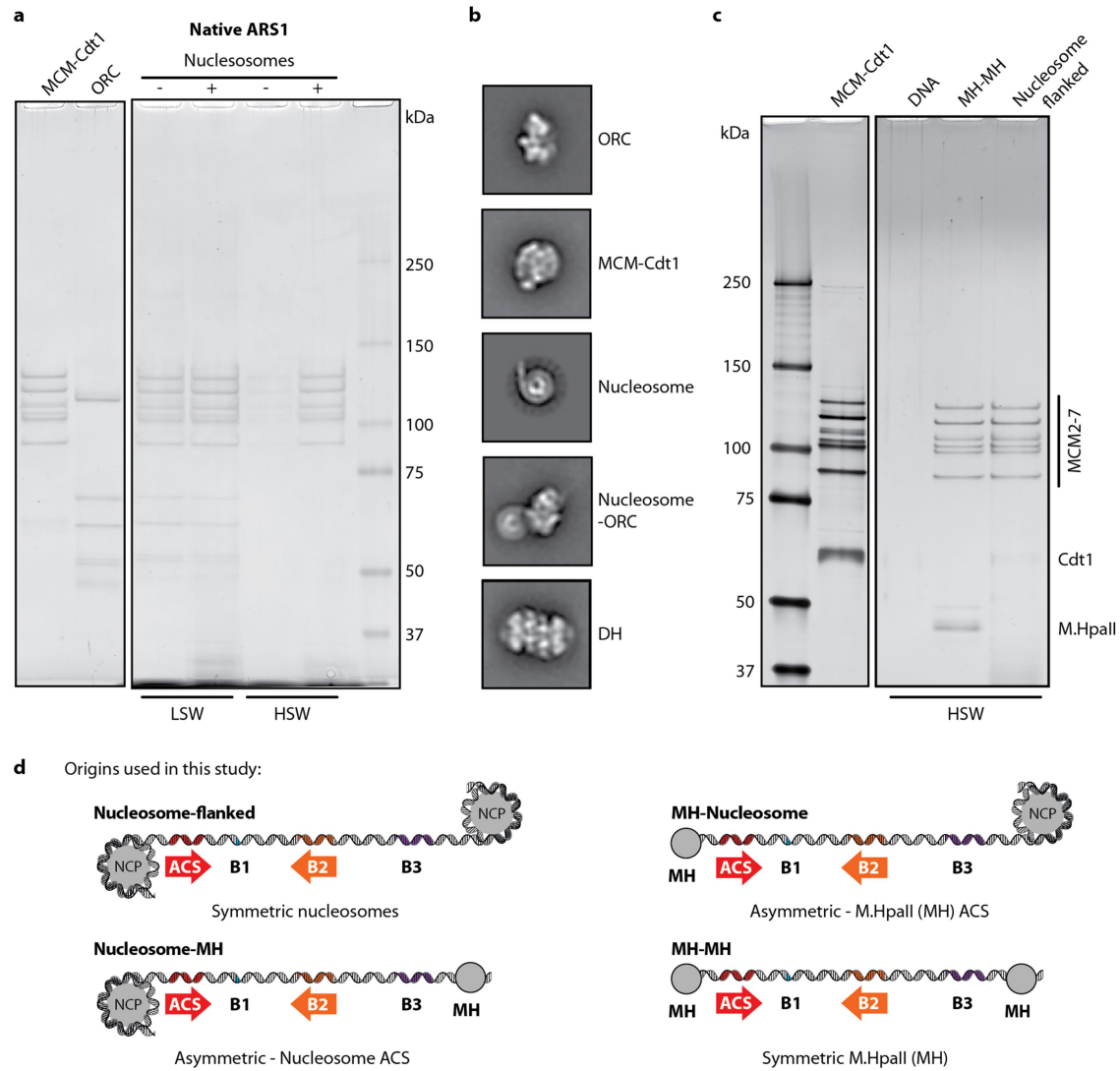
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1768-0>.

Correspondence and requests for materials should be addressed to A.C.

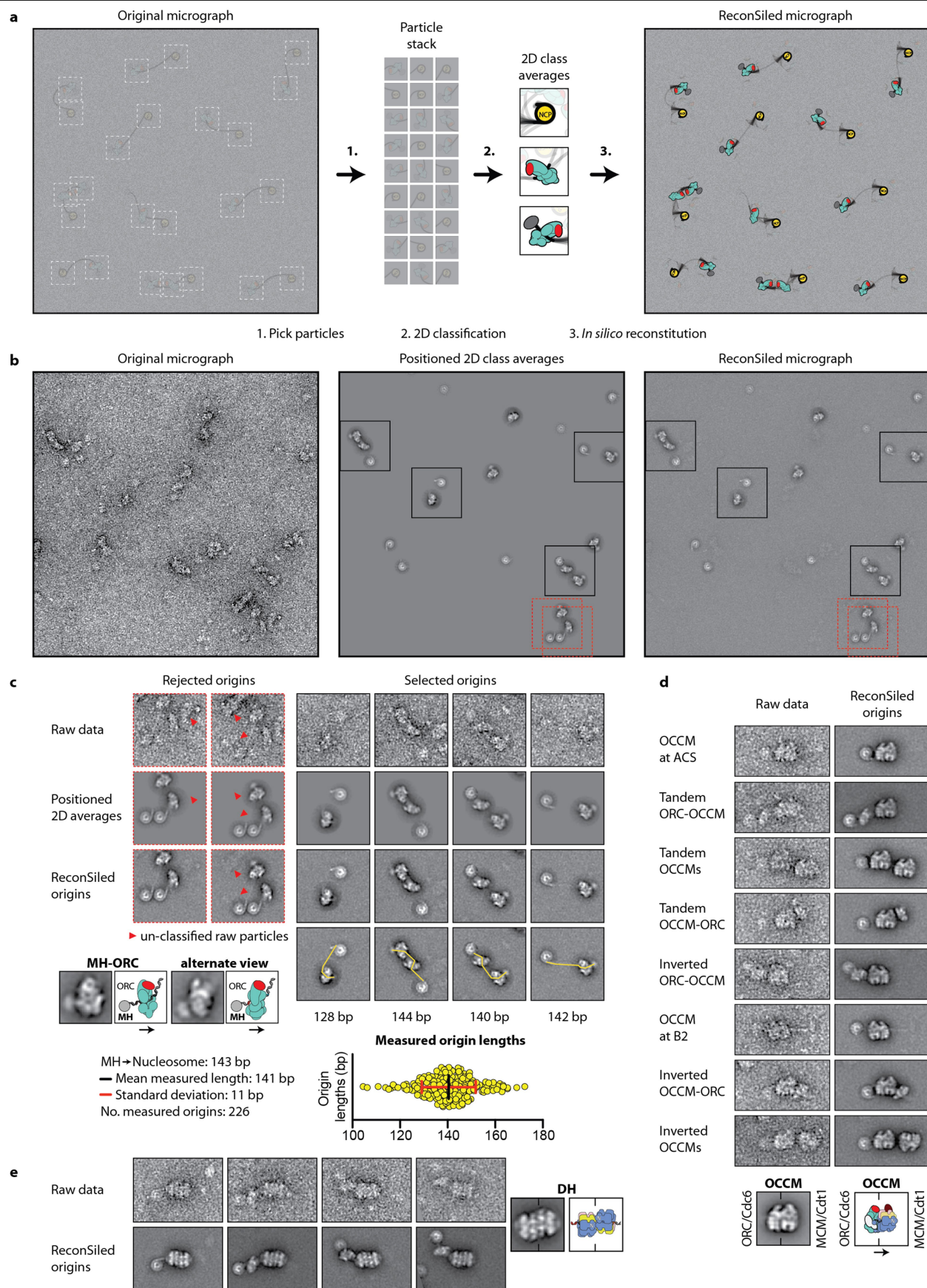
Peer review information *Nature* thanks Yuan He, Anthony Schwacha and Michael Trakselis for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Nucleosomes function as roadblocks that limit the linear diffusion of double hexamers. **a**, Loaded MCM particles are retained on short naked DNA when washed with low salt (300 mM NaOAc, LSW), whereas they slide off DNA when washed with high salt (500 mM NaCl, HSW). When chromatinized, the same DNA substrate retains MCM particles. In this experiment, soluble MCM loading reactions are bound to streptavidin-coated magnetic beads via a desthiobiotin tag on the origin DNA, before removing unbound proteins with high or low salt washes. **b**, Electron microscopy imaging of soluble MCM loading reactions yields 2D class averages of licensing complexes on chromatinized DNA. Despite the sample heterogeneity, recognizable classes can be obtained for ORC, MCM-Cdt1, nucleosomes, ORC that maps in close proximity to a nucleosome, as well as double hexamers.

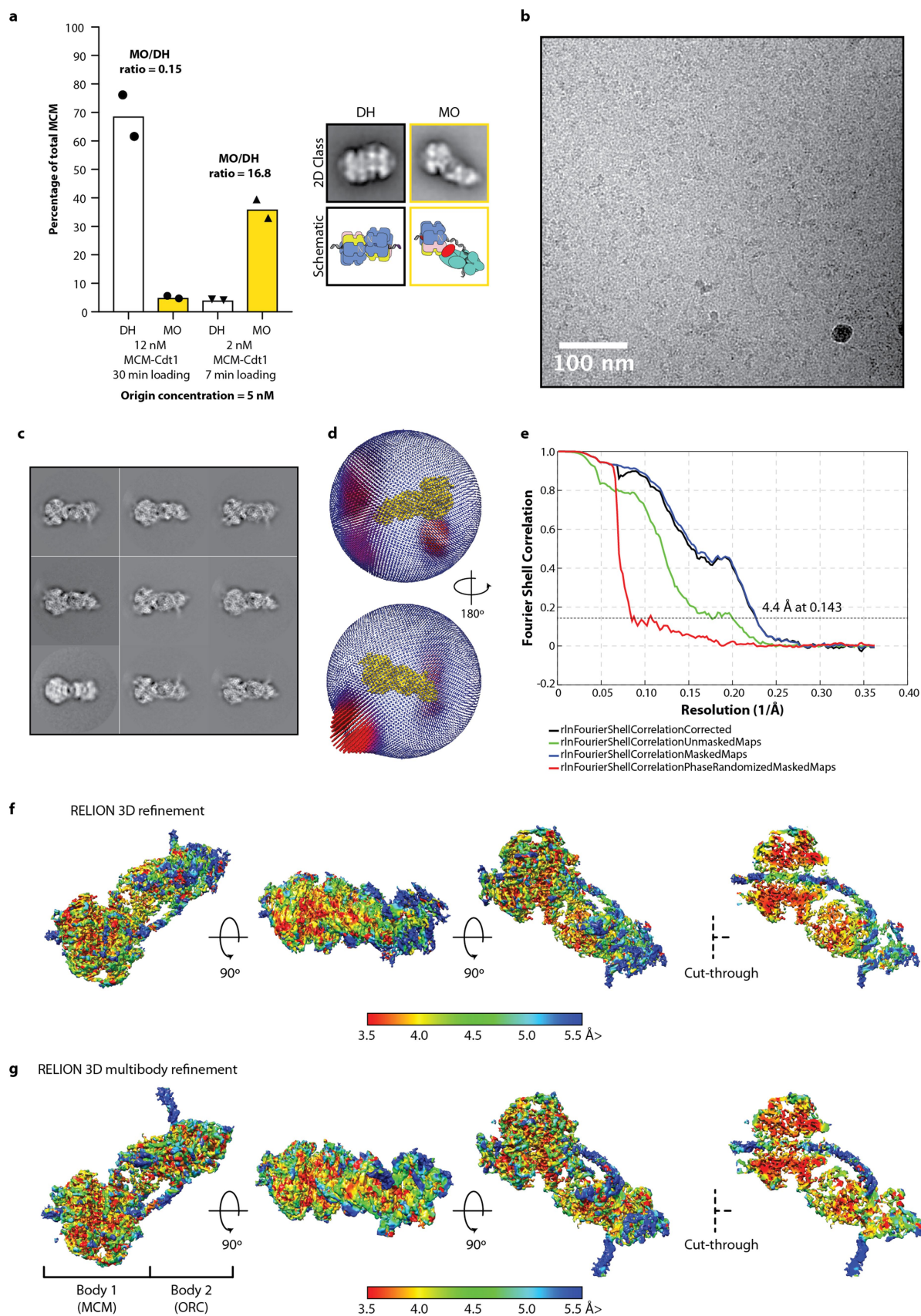
c, Comparison between MCM loading on chromatinized DNA and DNA containing M.HpaII (MH) roadblocks. After high-salt-wash treatment, equal amounts of loaded MCM helicases are retained on streptavidin beads, which indicates that nucleosomes are not required for efficient formation of double hexamers in this assay. **d**, Yeast replication origins centred on *ARS1*, containing the two inverted ORC binding sites, ACS (high affinity, red arrow) and B2 (low affinity, orange arrow). *ARS1* is flanked by nucleosomes, covalently attached methyltransferases (M.HpaII-M.HpaII) or a combination of the two to obtain asymmetric origins with recognizable features that mark the ends of the origin (M.HpaII-nucleosome and nucleosome-M.HpaII). For gel source data for **a** and **c**, see Supplementary Fig. 1.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | In silico reconstitution of origin licensing performed on asymmetric origins of replication. **a**, Cartoon depicting the ReconSil procedure, as performed to investigate the interactions between ORC and an asymmetric origin. Particles are picked on micrographs with a low signal-to-noise ratio. Two-dimensional averages are calculated. Averages are superposed to the raw micrographs, overlaid to the particles that contributed to their generation. For this purpose, particle coordinates are combined with alignment parameters derived from 2D classification. This approach yields a signal-enhanced view of single instances of molecular complexes bound to a flexible substrate (in this case, ORC binding to an entire origin of replication). **b**, Representative raw micrograph, 2D class averages positioned according to their constituent particles, and a micrograph of origins reconstituted in silico with positioned 2D class averages overlaid onto the original image. Instances boxed in black are selected, red are rejected. **c**, Left, origins might be rejected

owing to local particle clustering and aggregation, or because they contain visible raw particles that could not be classified (and therefore are not matched by a high-quality 2D average). This assay used M.HpaII-nucleosome origins that permit measurement of the length of origins because both the M.HpaII roadblock (next to ACS-bound ORC) and the nucleosome can be reconstituted. The measurement of origins reconstituted in silico was performed using ImageJ. **d**, Comparison of raw negative-stain electron microscopy data and origins reconstituted in silico for representative OCCM-bound origins shown in Fig. 1g. **e**, Example of origins reconstituted in silico (and corresponding raw images) showing double hexamers recruited to nucleosome-M.HpaII origins. ORC frequently rebinds to the ACS on origins that contain double hexamers, but shows no fixed interaction with the C-terminal face of the loaded double hexamer.



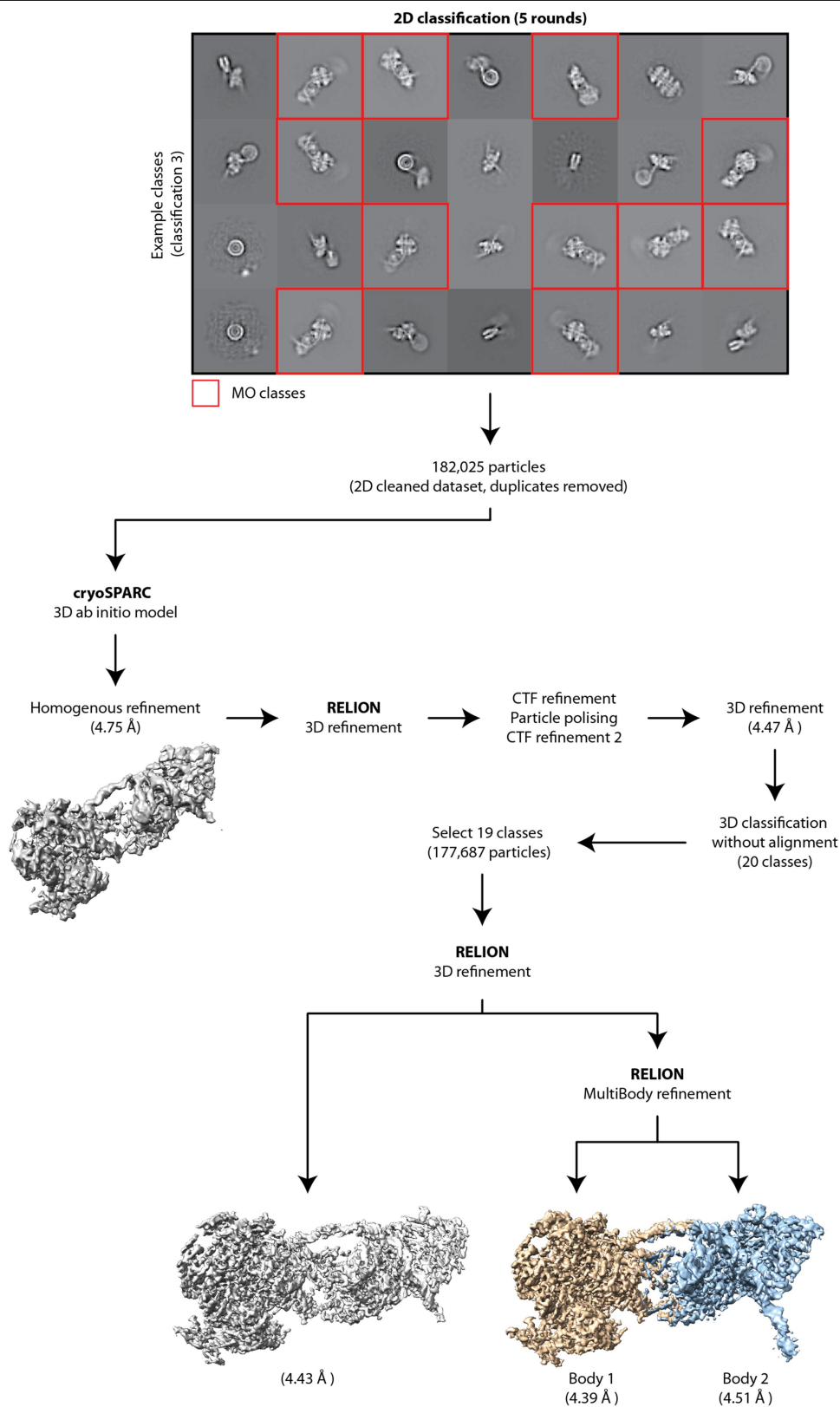
Extended Data Fig. 3 | See next page for caption.

Article

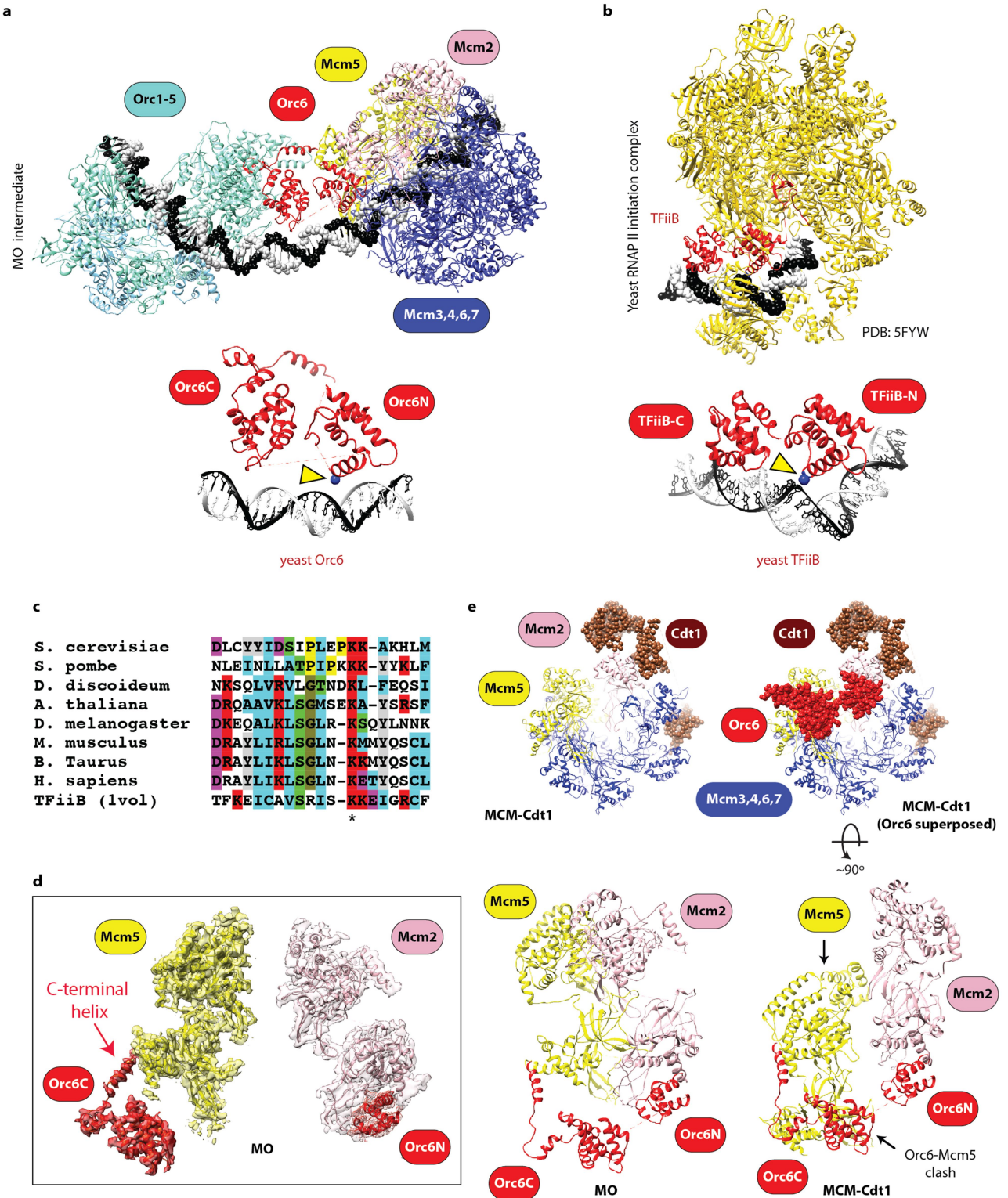
Extended Data Fig. 3 | Cryo-EM structure of the MO loading intermediate.

a, The MO intermediate is enriched when MCM–Cdt1 concentration is limiting, as quantified using negative-stain electron microscopy. An MCM loading reaction performed for 30 min in the presence of excess MCM–Cdt1 results in the majority of MCM helicases forming double hexamers on DNA. If MCM–Cdt1 concentration is limited and loading time is reduced (7 min), MO complexes form but do not mature into double hexamers; this indicates that the MO

intermediate is on the path to the formation of the double hexamer. Bar chart shows mean, $n = 2$ independent experiments. **b**, Example of an aligned movie. **c**, Resulting 2D averages. **d**, Angular distribution. **e**, Resolution estimated using gold-standard Fourier shell correlation. **f**, Three rotated views and a cut-through view of the MO 3D structure, colour-coded according to local resolution. **g**, Structure obtained by multi-body refinement, displayed as described for **e**.

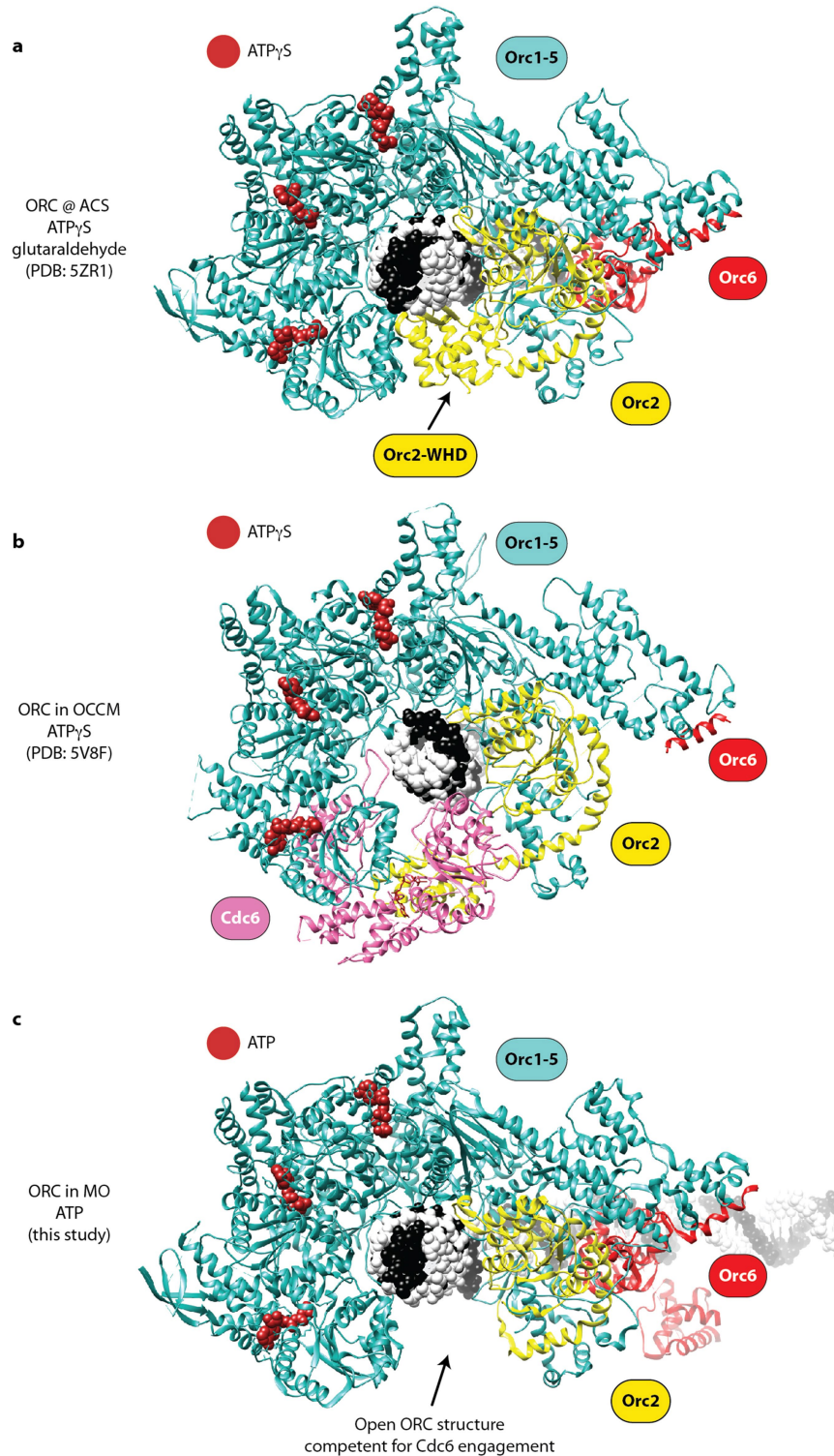


Extended Data Fig. 4 | Pipeline for generating the MO structure. Schematic shows the classification and refinement of the MO cryo-EM map.



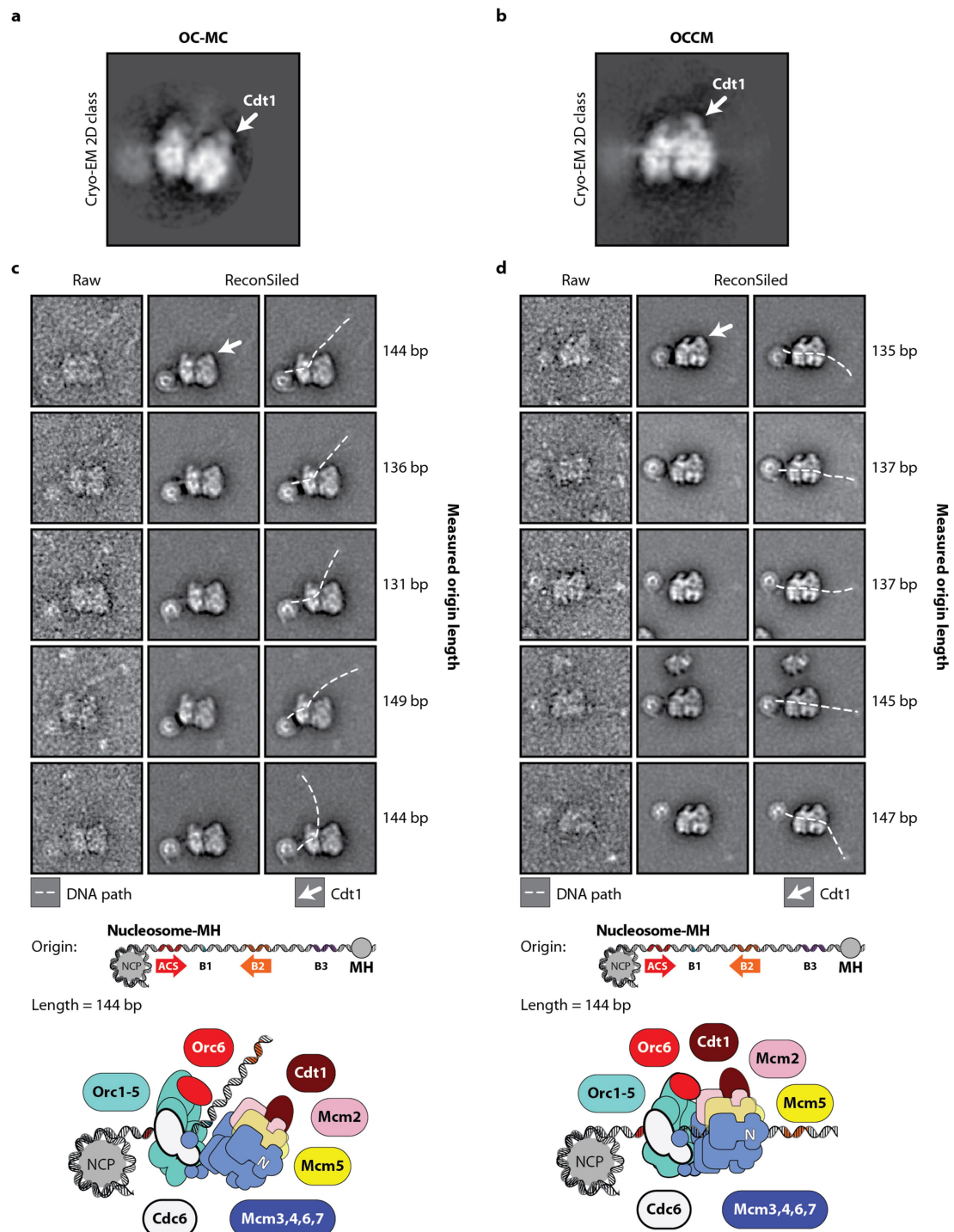
Extended Data Fig. 5 | A role for Orc6 in modulating MCM loading. **a**, Two elements connect ORC and the N-terminal face of MCM. One is Orc6 and the second is DNA, which is solvent-exposed between the ORC and MCM complexes owing to the bend induced by complex formation. **b**, Orc6 contains a domain architecture preserved in the related TFiiB transcription factor²⁹. Although the precise mode of DNA engagement for the N- and C-terminal domains of TFiiB and Orc6 differ, notable conservation can be detected. **c**, Sequence alignment between the N-terminal domain of TFiiB and the N-terminal domain of Orc6. The N-terminal domain of Orc6 contacts DNA through a conserved lysine that is also found in TFiiB. Mutation of the

equivalent lysine in *Drosophila* Orc6 affects DNA binding in vitro, as well as replication in extracts and cells³⁰. **d**, A conserved helix^{18,31} of the Orc6 C-terminal domain (Orc6C) touches the N-terminal helical bundle of Mcm5. The Orc6 N-terminal domain (Orc6N) touches the N-terminal helical bundle of Mcm2. Together, the N-terminal and C-terminal domains of Orc6 latch across the Mcm2-Mcm5 gate. **e**, No steric clash can be detected between Orc6 and Cdt1 when MO and MCM-Cdt1 are superposed via the N-terminal domain of Mcm2. However, the C-terminal domain of Orc6 severely clashes with the N-terminal domain of Mcm5 in this configuration. Only Orc6 from MO is shown in the MCM-Cdt1 superposed structure.



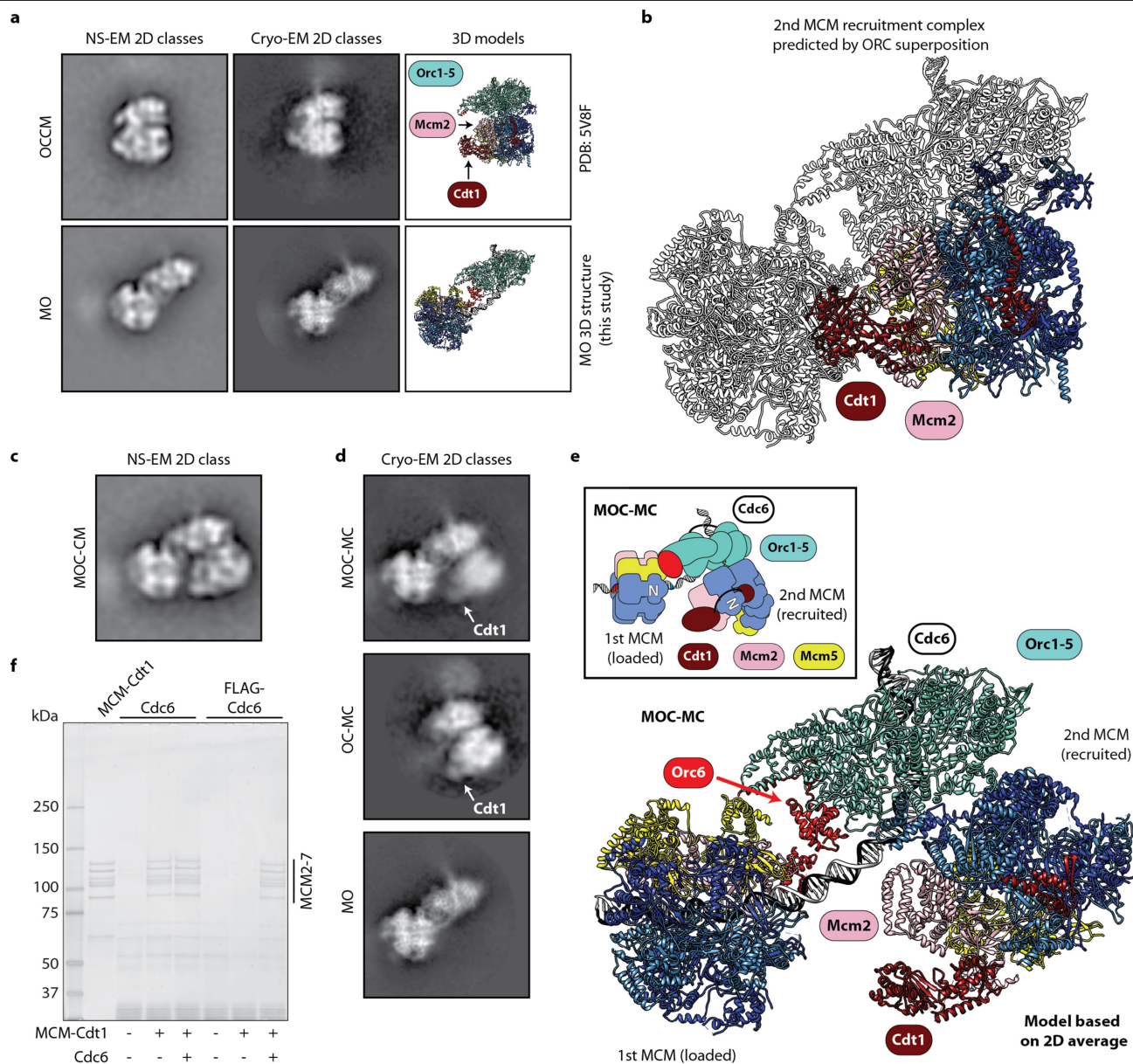
Extended Data Fig. 6 | Structure of ORC–DNA in different states. a–c. Comparison between the cross-linked ORC–DNA complex imaged in isolation (**a**), ORC–DNA in the OCCM complex (**b**) and ORC–DNA in the MO complex (**c**). Nucleotide occupancy appears the same in all three cases. It should be noted, however, that ORC–DNA alone and within the OCCM complex were co-incubated with ATP γ S, whereas ORC in MO was imaged in ATP. Orc2 in ORC–DNA contains a visible winged-helix domain (WHD) that topologically closes

ORC around DNA. ORC in OCCM is Cdc6-engaged. The Orc2 winged-helix domain is virtually absent in the cryo-EM map of the MO, which indicates that this domain is flexible. This discrepancy might reflect a different ORC configuration in MO, or the fact that the previously published ORC–DNA structure was stabilized by glutaraldehyde crosslinking. Despite Cdc6 being present in the sample, ORC in MO is not Cdc6-bound.



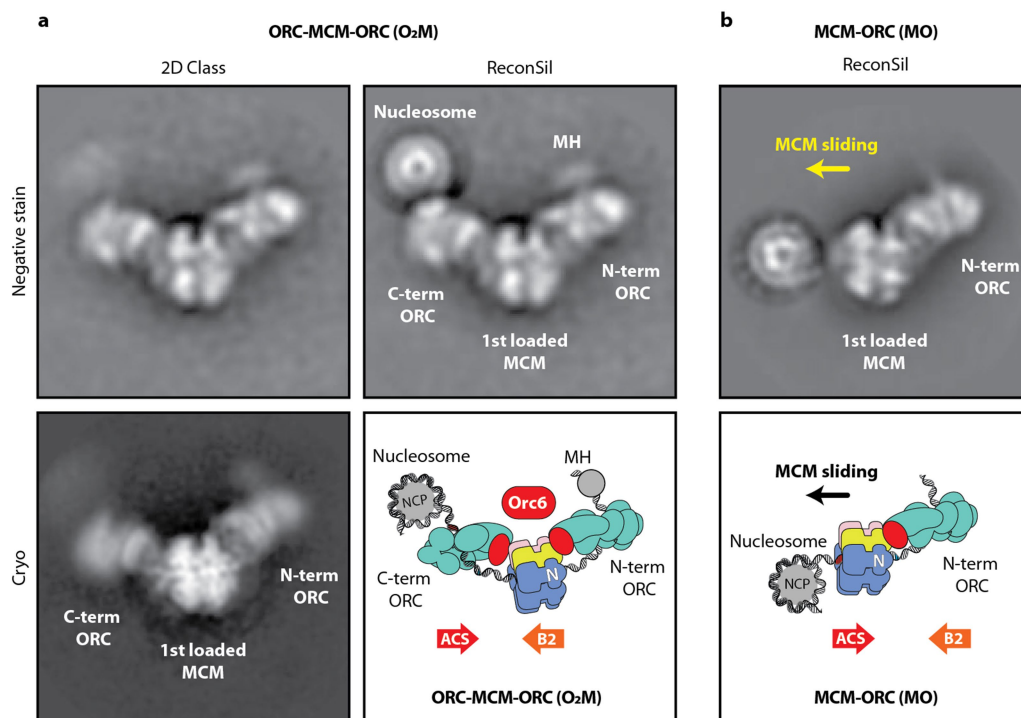
Extended Data Fig. 7 | OC-MC contains a recruited, but not DNA-engaged, MCM-Cdt1. a, b, Cryo-EM 2D class averages indicate that OC-MC is a pre-OCCM intermediate. **c, d**, This finding is confirmed by comparison of raw origins and origins reconstituted in silico, which permit visualization of the DNA path through OC-MC (**c**) and OCCM (**d**) in negative-stain experiments. In OC-MC, MCM-Cdt1 has engaged a DNA-bound ORC complex; however, DNA

remains outside the MCM channel. In this configuration, DNA is aligned to the Mcm2-Mcm5 gate, which can be located in the 2D images because of its proximity to the prominent N-terminal lobe of Cdt1 (white arrow). By contrast, DNA runs through the central channels of both ORC and MCM in the OCCM complex, in preparation for Cdt1 release and closure of the MCM ring.



Extended Data Fig. 8 | ORC in MO is perfectly positioned for loading the second MCM ring in the correct orientation for the formation of the double hexamer. a, Negative-stain and cryo-EM 2D classes, and 3D structures of OCCM (top) and MO (bottom), with the loading intermediates aligned via their respective ORC complexes. **b**, Three-dimensional model, based on **a**, of the proposed mechanism for recruitment of the second MCM. OCCM is shown superposed to the ORC of MO. This superposition places a second MCM-Cdt1 such that its Mcm2-Mcm5 gate is oriented for threading duplex DNA into the MCM channel. **c**, Negative-stain 2D class showing a post-MO loading intermediate, captured by supplementing MO complexes with MCM-Cdt1 before imaging. This class appears to be a second MCM recruitment complex, containing MO and an additional MCM-Cdt1. **d**, A cryo-EM 2D class average of

the post-MO complex (top) shows bent duplex DNA aligned to the Mcm2-Mcm5 DNA gate of the second MCM-Cdt1, captured before DNA threading. This is the same configuration that was previously identified for the OC-MC complex (middle). Alignment of the OC-MC and MO 2D classes by their respective ORC complexes matches the observed configuration of the second MCM recruitment complex, MOC-MC. **e**, Three-dimensional model of MOC-MC, based on the MCM-Cdt1 structure¹⁰, the MO structure (this study) and 2D class averages shown in **c** and **d**. **f**, Cdc6 is required for the loading of the second MCM helicase. Following immunodepletion of Flag-tagged Cdc6, MO is unable to load a second MCM; this results in a failure to form salt-stable double hexamers on DNA in the absence of additional Cdc6. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 9 | A second ORC can bind to a loaded MCM helicase before release of the first ORC. **a**, Negative-stain 2D class, nucleosome–M.HpaII origin reconstituted in silico, cryo-EM 2D class and schematic showing an intermediate on the path to the formation of the double hexamer. The intermediate contains a single-loaded MCM helicase (Cdt1 has been released) flanked by ORC at its C and N termini (ORC–MCM–ORC, or O_2M). The in silico reconstitution shows an entire origin, which spans a nucleosome, an ORC at the

C-terminal face of an MCM hexamer, an ORC at the N-terminal face of the MCM and a covalently linked M.HpaII. **b**, In silico reconstitution and schematic showing a Nucleosome–M.HpaII origin bound by MO. In this configuration the MCM in MO occupies the ACS site, which must have previously been bound by an ORC (as seen in the ORC–MCM–ORC complex in **a**). This observation demonstrates that MCM sliding towards the nucleosome has occurred.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	MO (EMDB-4980) (PDB 6RQC)	
Data collection and processing		
Magnification	105,000	
Voltage (kV)	300	
Electron exposure (e-/Å²)	50.4	
Defocus range (µm)	-2.7 to -4.2	
Pixel size (Å)	1.38	
Symmetry imposed	C1	
Initial particle images (no.)	6,287,507	
Final particle images (no.)	177,687	
Map resolution (Å)	4.4	
FSC threshold	0.143	
Map resolution range (Å)	3.3 – 6.3	
Refinement		Model Resolution
Initial model used (PDB code)	5ZR1 (ORC-DNA)	3.0
	6EYC (MCM-DH)	3.8
	6F0L (MCM-DH-DNA)	4.8
	5BK4 (MCM-DH-DNA)	3.9
FSC threshold	0.143	
Map sharpening B factor (Å²)	-149.415	
Model composition		
Non-hydrogen atoms	53,245	
Protein and DNA residues	6350	
Ligands (ATP, ADP, Mg²⁺, Zn²⁺)	15	
B factors (Å²)		
Protein	129.87	
Nucleotide	267.34	
Ligand	128.76	
R.m.s. deviations		
Bond lengths (Å)	0.009	
Bond angles (°)	1.411	
Validation		
MolProbity score	2.04	
Clashscore	8.98	
Poor rotamers (%)	0.54	
Ramachandran plot		
Favored (%)	89.5	
Allowed (%)	10.4	
Disallowed (%)	0.1	

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Gatan DigitalMicrograph and ThermoFisher EPU.

Data analysis

EMAN1 v1.9, EMAN2 v2.07, Xmipp v3.1, CrYOLO v1.40, MotionCor2, Gctf v1.18, RELION v3.04 and 2.1, CryoSPARC v0.6.5 and 2.5.0, UCSF Chimera 1.11.2, COOT v0.8.9.1, Phenix v1.14, MolProbity, ImageJ v1.50c, Tigris v0.3, HHPred

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Cryo-EM map and atomic model coordinates for the MO complex are deposited in the Electron Microscopy Data Bank and Protein Data Bank respectively under the accession code 6RQC and EMD-4980.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<p>Negative stain (NS) EM data were collected with the goal of obtaining sufficient particles to be able to perform 2D alignment and classification of the multiple species present in our biochemical reactions, including relatively rare, transient intermediates. Typically, 100-200 micrographs were collected. For our ReconSIL experiments, we aimed to obtain approximately 200 successfully reconstituted origins, which would enable us to identify multiple instances of any molecular events that occur at at least 1% origins.</p> <p>Cryo-EM data were collected with the goal of obtaining a high resolution 3D reconstruction of the MO complex, sufficient for the unambiguous assignment of the molecular components of the complex. Because our EM experiments imaged a mixed population of molecular components (ORC, Cdc6, MCM, Cdt1, DNA and nucleosomes) as they interact with one another to form DHs in the presence of ATP, we required a large data set to obtain sufficient MO particles for a high resolution reconstruction. In total ~22.5 k micrographs were collected from a single grid to obtain the structure presented in this paper.</p> <p>No statistical methods were used to predetermine sample size.</p>
Data exclusions	<p>For our ReconSIL experiments, samples were prepared with reduced concentrations to limit particle crowding and allow the clear identification of single origins of replication. Particles were picked and multiple rounds of 2D classification were performed to isolate particles contributing to the distinct molecular species in our samples. Picked particles that could not be aligned and classified were discarded and therefore were not reconstituted in silico. ReconSIL origins were evaluated and rejected if confident assignment of co-localisation to the same origin could not be made because either, i. the origin was in a region of clustered/aggregated particles or ii. If the origin contained additional particles that had not been 2D classified, and were therefore not overlaid with a 2D class average that would permit confident assignment of the molecular species.</p> <p>For other NS and cryo-EM experiments, micrographs were excluded for poor staining or ice contamination, respectively. Picked particles were aligned and classified and selected for downstream processing in 2D (NS and cryo) with particles that could not be aligned to a distinct class being excluded from further analysis. Cryo-EM data was further classified in 3D, with particles that negatively impact overall resolution of the final reconstruction being excluded.</p>
Replication	<p>All MCM-DH loading intermediates identified in this study (OC-MC, OCCM, O2M, MO and MOC-MC) were visualised in multiple experiments (both negative stain and cryo-EM). Multiple ReconSIL experiments investigating ORC binding and MCM-recruitment were also performed, yielding consistent results. The finding that truncation of the N-terminal domain of Orc6 permitted OCCM formation in the absence of ATP hydrolysis, but reduced the efficiency of both MO formation and DH loading in ATP was reproduced across multiple assays under different conditions (bead-based pull down assay; negative stain EM classification and quantification; MO assay; DH formation assay). The Cdc6 dependency of second hexamer recruitment to MO was confirmed in duplicate DNA-based pulldown assays visualised by SDS-PAGE and silver staining. The result was additionally confirmed by visualisation and quantification of MO and DH formation by negative stain EM.</p>
Randomization	<p>Randomization of samples is not relevant for a single particle electron microscopy study such as this.</p>
Blinding	<p>Blinding is not relevant for a single particle electron microscopy study such as this.</p>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	S.cerevisiae overexpression strains for DH loading factors were obtained from John Diffley and have previously been described in Frigola et al 2013.
Authentication	S.cerevisiae overexpression strains were checked for correct plasmid integration by PCR amplification from extracted genomic DNA.
Mycoplasma contamination	S.cerevisiae overexpression strains were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study

Author Correction: Forearc carbon sink reduces long-term volatile recycling into the mantle

<https://doi.org/10.1038/s41586-019-1756-4>

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1131-5>

Published online 24 April 2019; corrected online 28 June 2019

P. H. Barry, J. M. de Moor, D. Giovannelli, M. Schrenk, D. R. Hummer, T. Lopez, C. A. Pratt, Y. Alpizar Segura, A. Battaglia, P. Beaudry, G. Bini, M. Cascante, G. d'Errico, M. di Carlo, D. Fattorini, K. Fullerton, E. Gazel, G. González, S. A. Halldórsson, T. Ilanko, K. Iacovino, J. T. Kulongoski, E. Manini, M. Martínez, H. Miller, M. Nakagawa, S. Ono, S. Patwardhan, C. J. Ramírez, F. Regoli, F. Smedile, S. Turner, C. Vetriani, M. Yücel, C. J. Ballentine, T. P. Fischer, D. R. Hilton & K. G. Lloyd

In this Article, Tehnuka Ilanko should have been listed as an author, with the affiliation: Department of Geography, University of Sheffield, Sheffield, UK. She contributed to the gas compositional and C isotope analyses (see 'Author contributions'). The original Article has been corrected online.

Author Correction: Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity

<https://doi.org/10.1038/s41586-019-1755-5>

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1485-8>

Published online 21 August 2019.

Angel G. Solis, Piotr Bielecki, Holly R. Steach, Lokesh Sharma,
Christian C. D. Harman, Sanguk Yun, Marcel R. de Zoete,
James N. Warnock, S. D. Filip To, Autumn G. York, Matthias Mack,
Martin A. Schwartz, Charles. S. Dela Cruz, Noah W. Palm,
Ruaidhri Jackson & Richard A. Flavell

In the ‘Data availability’ section of this Article, the accession code ‘GSM133069’ should have read ‘GSE133069’. The Article has been corrected online.

Author Correction: Metastatic-niche labelling reveals parenchymal cells with stem features

<https://doi.org/10.1038/s41586-019-1697-y>

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1487-6>

Published online 28 August 2019.

Luigi Ombrato, Emma Nolan, Ivana Kurelac, Antranik Mavousian, Victoria Louise Bridgeman, Ivonne Heinze, Probir Chakravarty, Stuart Horswell, Estela Gonzalez-Gualda, Giulia Maticchione, Anne Weston, Joanna Kirkpatrick, Ehab Husain, Valerie Speirs, Lucy Collinson, Alessandro Ori, Joo-Hyeon Lee & Ilaria Malanchi

In this Article, Supplementary Data file 3, containing the sequence for sLP-mCherry, was missing. This file has now been added to the original Article, and a citation has been included at the end of the first sentence of the ‘Labelling system’ section of the Methods.

In addition, in the ‘Data availability’ section, the Gene Expression Omnibus accession number for the single-cell RNA-sequencing datasets was incorrectly listed as ‘GEO13150’ instead of ‘GSE131508’, and the link for the Proteomics Identifications Database accession ‘PXD010597’ was incorrect. These errors have been corrected in the original Article online.



CHINA'S PHD STUDENTS GIVE THEIR REASONS FOR MISERY

The nation's junior scientists are struggling for work–life balance, careers guidance and emotional support. **By Chris Woolston and Sarah O'Meara**

PhD students in China face outsize challenges as they try to complete their degrees, according to *Nature's* fifth biennial survey of PhD students.

On many measures, students in China fare worse than students in other parts of the world. One telling number: only 55% of the Chinese students who responded to the survey said that they were at least partially satisfied with their PhD experience. For the 5,630 respondents outside China, the satisfaction rate was 72% (see 'A nation apart').

The self-selecting survey was translated into Chinese as part of an effort to increase participation inside the country; it was created with Shift Learning, a market-research company based in London, and the full data set is available at go.nature.com/2nqjndw. The outreach paid off, with responses from 690 students in China – the highest response in the survey's 8 years. Through survey answers and free-text comments, the students expressed a relatively troubled view of PhD life marked by pockets of optimism and resilience.

Some respondents used the survey's comments section to point out the positives of PhD programmes. One student wrote that, compared with other sectors of Chinese society, such as politics and industry, the academic system encourages "freedom, creativity, discovery, and a greater acceptance of unexpected failure". Another said that the system is "relatively free and fair" and that PhD students are "able to do the things they like based on their own interests". One respondent singled out the opportunity for

“independence and innovation”. And one reported being satisfied “overall”, but added that there is “much room for improvement”.

Most respondents who felt prompted to comment expressed a more negative outlook. “Do not do a PhD in this country,” one student wrote. “No one will help you. No one will understand you. This is a prison.” Another wrote: “PhD pressure is too great, beyond my expectations.” In many ways, pressure is built into the system, says Di Chen, a cell biologist at Nanjing University. “Graduate students from most institutions are required to have at least one first-authored paper with certain levels of impact factor to get their PhD degree,” he says. “Therefore, everyone has to be productive, which is nearly impossible.”

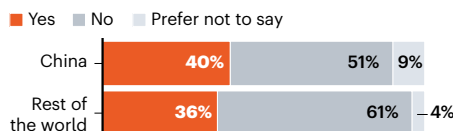
The PhD ranks are becoming crowded. According to the Chinese Ministry of Education, 95,502 new PhD students enrolled in 2018, bringing the total PhD-student population to 389,518. By comparison, little more than 70,000 new PhD students enrolled in 2013, and fewer than 62,000 enrolled in 2009. Some observers think that the supply of PhD students is greater than the nation’s current educational system or job market can completely support. “The whole infrastructure needs to be reformed,” Chen says. “I personally believe reducing the number of PhD students might help.”

PhD programmes remain popular, but the survey found that regrets are widespread. Asked what they would do differently, 22% of respondents said that they would change their supervisor, 36% that they would change their area of study, and 7% that they wouldn’t pursue a PhD at all. Forty-five per cent of respondents said that their programme fell short of expectations. Outside China, that proportion fell to 36%. At the other end of the spectrum, just 5% of Chinese respondents said that their PhD programme exceeded expectations; the corresponding rate in the rest of the world was more than twice as high.

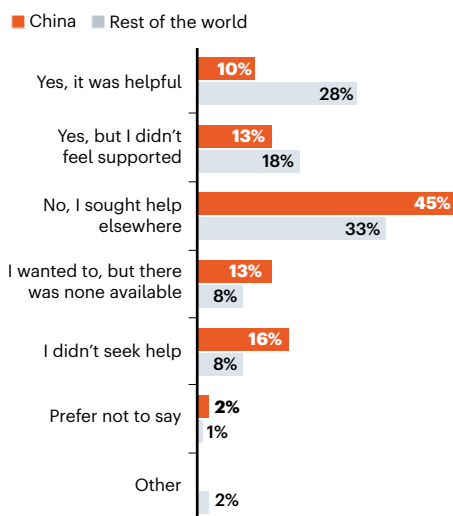
EMOTIONAL TOLL

Like PhD students everywhere, students in China face strains that can threaten mental health. Relatively few Chinese students found help at their institutions.

Q. Have you ever sought help for anxiety or depression caused by PhD study?



Q. Did you seek help for anxiety or depression within your institution?



Percentages do not all add up to 100 because of rounding.

Often, the expectations might have been too lofty from the beginning, says Qilin Zhou, a chemist at Nankai University in Tianjin. “Many students think scientific research is beautiful and romantic before entering the laboratory,” he says. “When they start to do research, they will inevitably encounter various difficulties.”

Nancy Li, a master’s student who dropped out of a PhD pharmacy programme at a leading university in China, isn’t surprised that so many Chinese students struggle with the reality of

a PhD programme. “A considerable amount of Chinese PhD students are not adequately prepared for PhD study and are in need of more guidance, including career advice and also psychological counselling,” she says.

Heavy tolls

Comments and survey answers put a spotlight on the emotional toll of PhD work. One student wrote that she wished she had known “how a PhD would affect my mental health and work–life balance”. She wasn’t alone. In the survey, 40% of respondents from China said that they had sought help for depression or anxiety caused by their PhD programme (see ‘Emotional toll’). That’s slightly more than the 36% of respondents in other parts of the world who sought help. For students in China, support is unlikely to be close at hand. Of those who sought help, only 10% said that they had benefited from assistance at their home institution. In other parts of the world, that figure was 28%.

In a positive development, students in China were less likely than were students elsewhere to complain about mistreatment. Only 15% reported bullying, compared with 22% in the rest of the world. Likewise, the percentage who reported discrimination or harassment (12%) compares favourably with the rate for respondents from other nations (22%).

Anxiety can come from many directions. For one thing, Chinese students face many demands on their time, although in lower numbers than elsewhere. More than half of all respondents (53%) reported working more than 40 hours a week. In the rest of the world, 79% of students reported putting in such hours; it could be that China’s percentage is smaller because the proportion of part-time students is higher. Fifty-four per cent agreed with the statement that “there is a long-hours culture at my university, including occasionally working through the night”. In China, as elsewhere, those long hours in the lab come with consequences: 45 per cent of respondents in China said that they were dissatisfied with their work–life balance. In the rest of the world, 38% shared that complaint.

Chong Tian, a chemist at the University of Manchester, UK, says that she regularly put in long hours – up to 11 hours a day, 6 days a week – during her PhD programme at Tsinghua University in Beijing. She says that she didn’t complain. “Working overtime is a common phenomenon in the whole society,” she says. “I enjoyed my project and always pushed myself to work harder to get results ASAP.”

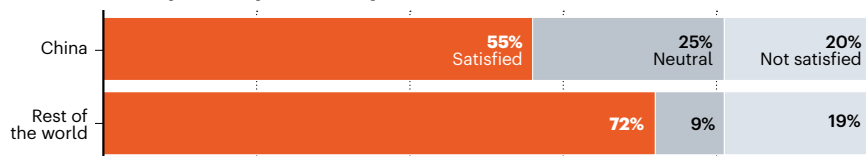
Uncertain prospects

Like their counterparts in other countries, PhD students in China also worry about job prospects after graduation. Nearly 90% of students ranked uncertainty about careers as one of their top-five concerns. On a more optimistic note, 70% of respondents think

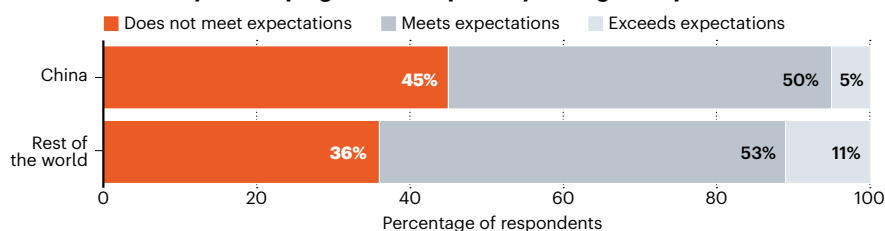
A NATION APART

Compared with their peers in the rest of the world, PhD students in China are less likely to find satisfaction – and more likely to find disappointment – in their programmes.

Q. How satisfied are you with your PhD experience?



Q. To what extent does your PhD programme compare to your original expectations?



that their PhD work will “substantially” or “dramatically” improve their job prospects, putting them slightly ahead of their peers in the rest of the world on that score (see ‘High hopes’). Despite a tight and competitive market for jobs in academia, the dream of a university research position remains powerful. Nearly 70% of respondents said they would most like to work in academia after graduation. By comparison, 55% of respondents outside China shared that goal.

Some worries are warranted. A nationwide survey found that 83% of new PhD recipients were employed in 2017, putting them slightly behind those with master’s degrees (85%) and vocational degrees (89%).

Solid careers advice could ease concerns about the future, but that advice isn’t always available. Nearly half of the respondents in China said that they had reached their career decisions on the basis of their own research, and another 28% credited family influence. Just 29% said they had based their decisions on advice from their supervisor. Overall, 46% of respondents said that they were dissatisfied with their careers guidance, putting them on a par with students elsewhere.

As a rule, the survey found, students in China are given little time to speak to their supervisors or principal investigators (PIs) about their careers, or about anything else. The majority, 52%, reported spending less than an hour one-to-one with their supervisor each week. Outside China, that figure was 49%. “Unfortunately, many supervisors do not provide enough help and guidance to students because they are busy applying for grants and other business,” Zhou says.

Several respondents complained that their lab felt more like a business operation than a training ground. As one put it, “the PI has all of the power. Everyone else in the lab is just a factory worker.” Li says that many labs have time clocks that record when each member arrives and leaves. “It is not so much a teacher–student

“The PI has all of the power. Everyone else in the lab is just a factory worker.”

relationship as it is an employer–employee relationship,” she says.

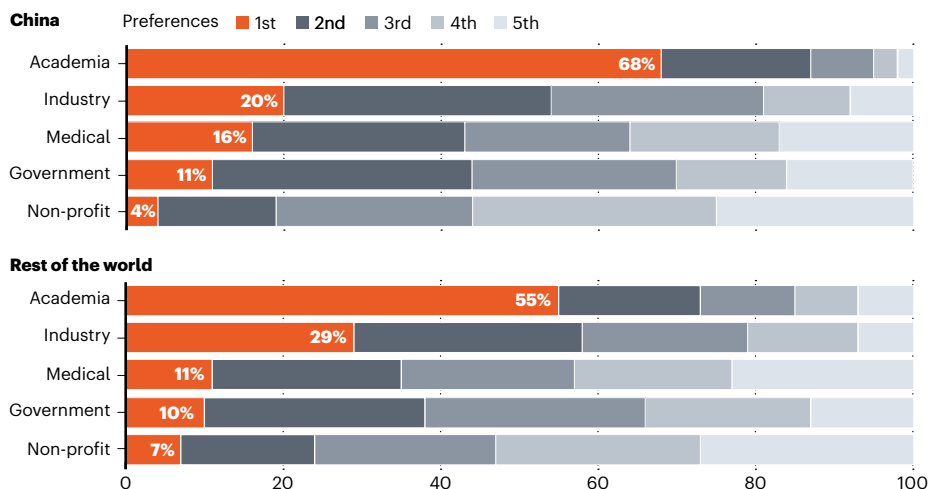
Chen notes that PIs themselves are under pressure. “PIs are evaluated mainly based on publications, especially those from prestigious journals, which sometimes require very labour-intensive experiments,” he says. “Compared to PIs from other countries, Chinese PIs seem to have more time-consuming duties from administration, lab management, family and so on.”

Zhou says that a growing number of Chinese academics are realizing that they have an

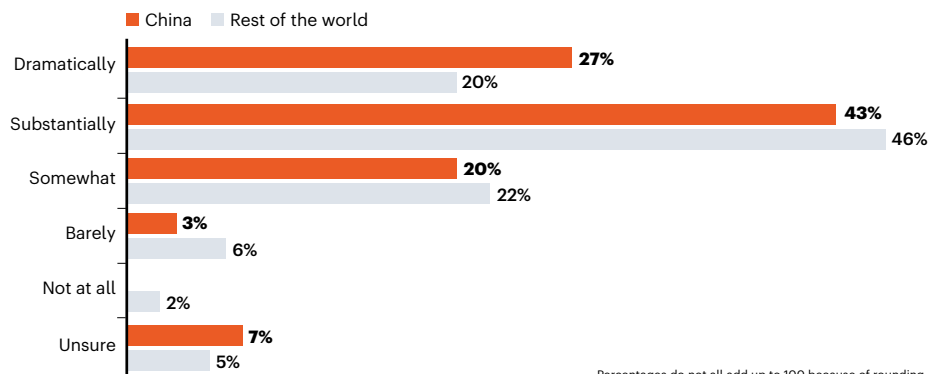
HIGH HOPES

PhD students in China are holding on to dreams of academic careers, and they’re relatively hopeful that a degree will boost their chances in the job market.

Q. Which of the following sectors would you most like to work in (beyond a postdoc) when you complete your degree? (Showing first choices only)



Q. How much do you expect your PhD to improve your job prospects?



Percentages do not all add up to 100 because of rounding.

obligation to their teams. “Things are already changing for the better, albeit slowly,” he says. “Supervisors should spend more time with students to give them help and guidance.”

Students could be doing more to get the advice they need, says a materials scientist at the Institute of Advanced Materials and Technology in Beijing, who prefers to remain anonymous. “I always tell my students that the speed of their growth depends strongly on how often they meet me,” the scientist says. “Like most Chinese professors, I’m busy, of course – but I always have time, or can find another time, if they want to discuss anything with me.”

PhD perseverance

Despite all of the challenges, respondents to the survey still found things to like about their graduate training. Asked what they enjoyed most about life as a PhD student, 27% singled out the university/academic environment – making that the most popular response. Others cited the intellectual challenge, the opportunity to be creative and the chance to work with bright and interesting people.

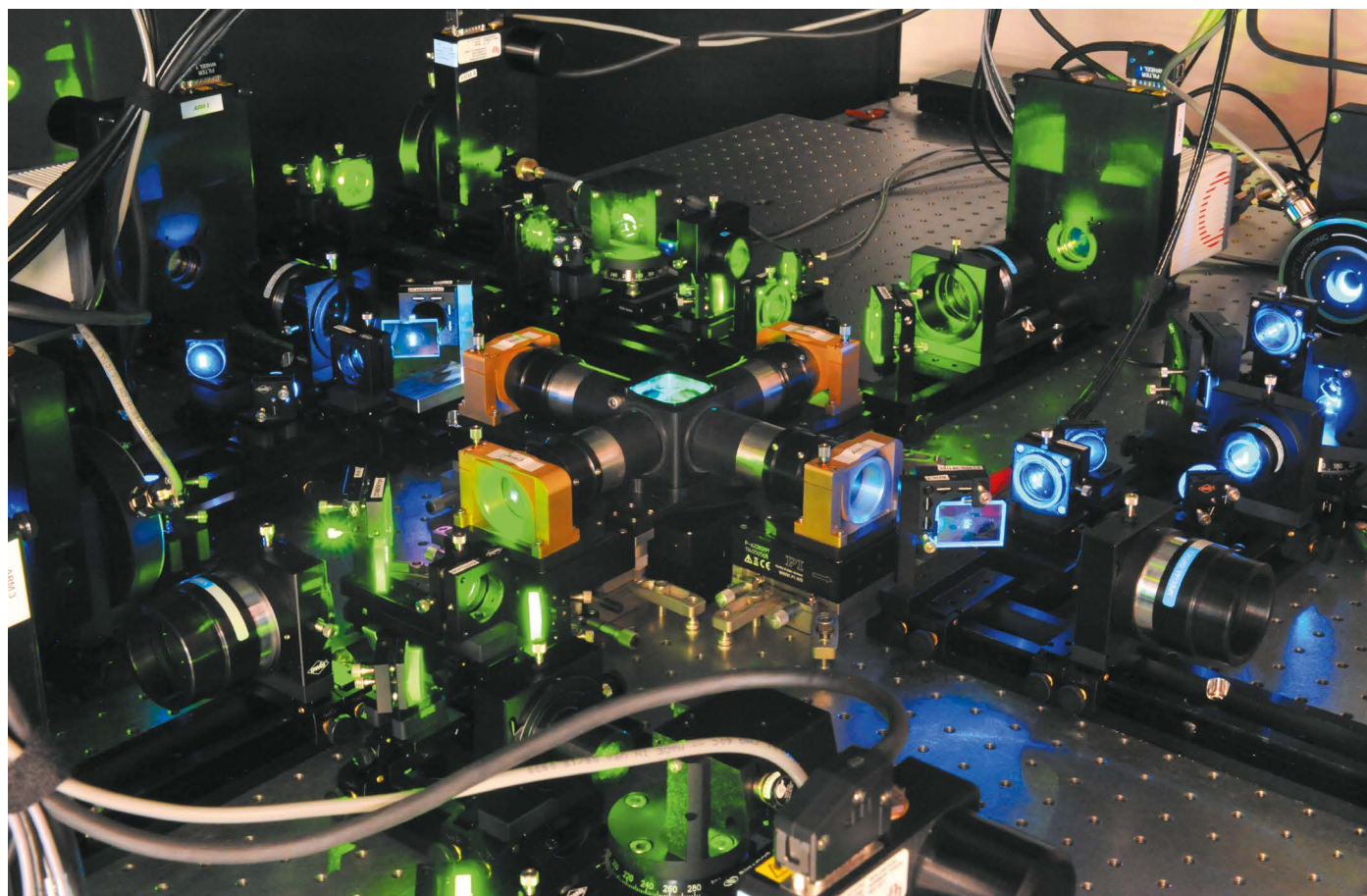
When asked if they were satisfied with their decision to pursue a PhD, 62% of respondents

said yes. That’s significantly behind the 76% of respondents from the rest of the world who expressed satisfaction. But Chinese academia is still in its relative infancy – just 18 PhD students were enrolled in PhD programmes in 1978 – and there are reasons to suppose that it will become more effective, rewarding and satisfying for students. “Science is developing fast in China,” Zhou says. “As the economy develops, more students will want to stay in academia.”

Looking back, Tian says, she values the training she received at Tsinghua University. She says that her adviser was always around and available, and she got useful careers advice from her “chemistry brothers”. She thinks she’s on the right track, but adds that time will tell. “No one can be adequately prepared for the future because we never know what will happen in the next second,” she says. “The most important thing is to figure out what we want, to be brave, and to try.”

Chris Woolston is a freelance writer in Billings, Montana. **Sarah O’Meara** is a freelance writer in London, UK.

Additional reporting by Kevin Schoenmakers.



Light-sheet microscopes use low-intensity lasers to study live tissue for extended periods.

GO BIG OR GO HOME

Microscopy is turning into ‘mesoscopy’ as researchers set their sights on ever-larger biological samples. **By Jeffrey M. Perkel**

In a sunny third-floor office at the Howard Hughes Medical Institute’s Janelia Research Campus in Ashburn, Virginia, Philipp Keller is showing off the optics for his latest microscope. They’re not much to look at – yet.

“What we have right now in the building is this,” Keller says, pointing to his computer screen. “Basically, these are slabs of glass as they would be drawn out of a continuous-flow furnace.”

The slabs resemble book-sized blocks of ice, standing as if on a library shelf. Keller, a physicist, hopes to use them to build a new type of microscope, one that can achieve high resolution and that can handle specimens of a size that has long been on biologists’ wish list.

In biological microscopy, he says, researchers can either look at big samples at low resolution, or small samples at high resolution. It hasn’t really been possible to look at big samples – larger than about one cubic millimetre

– and pick out cellular or even finer details.

“It’s usually a trade-off,” he says. “You can either get a macroscopic view, or you can get a high-resolution, zoomed-in view. And then the only way to combine them both is that you do some kind of massive tiling strategy where you take a tiny imaging volume and you just raster it through the entire sample in 3D” – a time-consuming and computationally demanding process.

Increasingly, researchers are developing ‘mesoscopes’ – mesoscale microscopes – to circumvent that challenge. These instruments can capture cellular and even subcellular processes across samples that can exceed one centimetre in size. The resulting data sets provide an unprecedented perspective. As Gail McConnell, an optical physicist at the University of Strathclyde in Glasgow, UK, who developed one such mesoscope, puts it: “It’s almost macro photography, but with higher resolution.”

Biological microscopy is all about compromise. To follow a fast-moving process, for instance, researchers typically capture many images in quick succession, with short exposure times. The sample must therefore be very bright to provide as many photons as possible in the time available. That requires more input light energy, which can kill (or at least bleach) the sample. As a result, such imaging usually cannot be done for long.

Big objectives

Similarly, systems that can capture fine cellular detail tend to have a narrow field of view. Point-scanning confocal microscopes produce sharp images of subcellular structures by scanning a tightly focused laser beam across a sample, exciting fluorescence in the sample pixel by pixel.

Such imaging is “fabulous for very tiny tissue volumes”, McConnell says, but it cannot be applied to large specimens – such as

late-stage mouse embryos – “because of the low numerical aperture of a low-magnification lens” that would be needed for such large samples. Numerical aperture (NA) refers to the ability of a lens to capture light; a higher NA usually corresponds to higher magnification and a shorter working distance between the objective lens and the sample.

To overcome that problem, McConnell teamed up with confocal-microscopy developer Brad Amos, also at Strathclyde, to build a macro-scale objective lens with the unusual combination of a high NA and low magnification, capable of providing both a wide field of view and high spatial detail. The result is the ‘Mesolens’, a custom optic that can image over a 6-millimetre-wide field of view with 0.7-micrometre lateral and 5-micrometre axial resolution, and a 3-mm working distance – sufficient to distinguish objects that are about one-tenth the diameter of a typical mammalian cell¹.

The Mesolens looks like an objective lens on steroids: “I would liken it to roughly the same length and width as an adult human arm,” McConnell says. It took a decade to build, and its sheer size – its glass elements are nearly three times the diameter of the typical microscope objective lens – makes it incompatible with many off-the-shelf components, McConnell notes. “The tolerance with which they have to be ground and polished” – not to mention aligned relative to one another – “becomes much more stringent” than with conventional lenses, she says.

Suitable detectors were a problem, too. One reason it took McConnell and Amos so long to build the Mesolens was that they had to wait for wide-field sensors that could capture the resulting photons, she says. Amos co-founded a spin-off company to commercialize the Mesolens, but “we’re not really in a commercial space at the moment”, says McConnell (who holds no stake in the company). “We are working on a lens prescription that will hopefully be easier to make.”

Still, McConnell’s team has used its design to begin to address biological questions, including the architecture of bacterial biofilms, which are collections of microbes that grow on a surface inside a film that they secrete, and which are often associated with disease. “We’re seeing new and emergent properties of these biofilms that could potentially inform our knowledge about tackling antimicrobial resistance,” she says.

Qionghai Dai, an information scientist at Tsinghua University in Beijing, has also been tackling the problem of imaging centimetre-scale samples. Recognizing a huge gap between high-resolution microscopy and macro-scale techniques such as functional magnetic resonance imaging and computed tomography, Dai’s team set about developing a gigapixel microscope with high

spatio-temporal resolution, says Lingjie Kong, a physicist in the lab.

The result of that work is RUSH, an instrument that features a custom objective lens, a curved array of 35 cameras and a computational system that can capture and analyse data in real time. Whereas the Mesolens can produce some 4 million pixels’ worth of data per second, Kong says, RUSH produces 5.1 billion, imaging a $10 \times 12 \text{ mm}^2$ field of view at 30 frames per second. That’s enough to image the entire surface of the mouse brain in a single shot, and to resolve individual sub-cellular organelles called mitochondria².

Dai’s team used its system to track fluorescently labelled blood cells travelling across the surface of a mouse brain, and to monitor neural activity in freshly prepared human-brain slices. Studies in non-human primates are being planned. A second-generation RUSH, featuring higher resolution and data throughput, is in development, Kong says, as is a commercial version of the instrument. “It is expected to be available by next year.”

Mirror mirror

To tackle its big-sample challenges, Keller’s team took a cue from astronomy. Large telescopes use mirrors because they are lighter and easier to fabricate than lenses. Mirrors also lack many of the aberrations that can degrade performance as lenses get larger. So why has nobody used them to build a microscope? “To be honest, I don’t know exactly why they haven’t done it,” Keller concedes.

“It’s almost macro photography, but with higher resolution.”

His team is no stranger to custom microscopes. Over the past decade, it has developed a succession of increasingly complex ‘light-sheet’ systems, in which a plane of low-intensity laser light is projected through the sample. Images are captured from the side (that is, at 90 degrees to the plane of light) to maximize imaging time and minimize photodamage. These microscopes, assembled on optical workbenches that dominate the small rooms in Keller’s lab, can record the cellular dynamics and neural activity of developing fruit flies and zebrafish for hours at a time.

In October 2018, Kate McDole, a research scientist in Keller’s group, described a combination microscope and environmental-chamber design that extended that period to two days. She has used it on a much more challenging subject: in tandem with a powerful suite of custom analysis software, she was able to image and track every cell in a mouse embryo over 48 hours. During that time, the embryo grows 250-fold in volume³.

Custom microscopes provide extraordinary

flexibility, but aren’t necessarily easy to use. McDole estimates that only seven of ten experiments run to completion. Sometimes the microscope misbehaves, and sometimes the sample does. Sometimes the carbon dioxide tank, required for maintaining the animal’s growth, gets shut off, and sometimes the microscope’s computer system decides to update during an experiment – a not-infrequent occurrence, she says. “Murphy” – as in Murphy’s Law – “follows me around like a puppy,” McDole quips.

The key element to the lab’s in-development ‘mirror microscope’ project, led by optical engineer Dan Flickinger and postdoc Benquan Wang, is a concave mirror 168 mm in length, which Keller says cost about US\$6,000 to make. With a $14 \times 14 \text{ mm}^2$ field of view, a high NA of 1.0 and a detection array comprising a dozen cameras, the system should capture some 200 times more of the sample than most commercial microscopes, yet still resolve sub-cellular detail, Keller says.

Working with another team at Janelia, he hopes to use the mirror-based system to capture the neural connectivity, or ‘connectome’, of an entire mouse brain. And he hopes to record the brain activity of several zebrafish simultaneously as they interact and swim in a small dish. Data sets could run to petabytes, Keller says – equivalent to what his team currently generates in a year. McDole’s first mouse data set was 20 terabytes, she says, “and it broke every piece of imaging software that we had”.

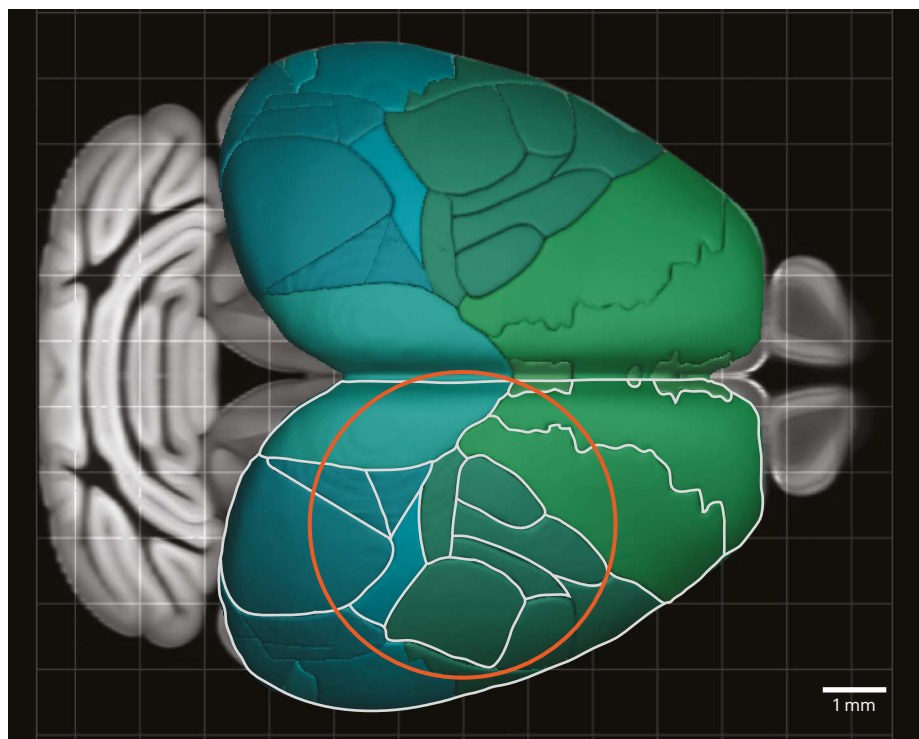
First, however, the team will have to show that the design works. Preliminary testing is set to kick off this month. “I am confident that by the end of the year we will have the first image that has been formed with an entirely mirror-based detection system,” Keller says.

Two photons are better than one

Light-sheet microscopy works best for relatively transparent samples. More opaque specimens call for other techniques, and researchers are making progress in applying them to large fields of view, too.

In 2016, teams led by microscopy developers Fritjof Helmchen at the University of Zurich in Switzerland, Spencer Smith, then at the University of North Carolina, Chapel Hill, and Karel Svoboda at Janelia independently described innovative ‘two-photon’ microscopes that can image macro-scale samples^{4–6}. These devices use ultrafast, long-wavelength laser pulses that excite fluorescence in the sample, but only in a sharply defined plane. The resulting images are high-contrast up to 1 mm deep, but cover only about 1 mm^2 . Svoboda’s system, called the 2p-RAM mesoscope, provides a $5 \times 5 \text{ mm}^2$ field of view – enough to take in the entire surface of the mouse brain. Helmchen’s and Smith’s systems image $1.8 \times 1.8 \text{ mm}^2$ and more than 9.5 mm^2 , respectively.

When going big, Svoboda says, it’s not



The 2p-RAM instrument's field of view (orange) can access much of the mouse brain at once.

enough to super-size the objective; multiple parts must be reengineered to accommodate how light propagates through such optics. In fact, the reason his team settled on a $5 \times 5 \text{ mm}^2$ field of view, he says, is that larger detectors weren't available: "5 mm is right now the limit, unless you want to throw away a lot of signal. And in this business, we hate to do that," he says.

Such systems allow neuroscientists to see the forest for the trees. Trying to decode neural communication by studying one brain region at a time, Svoboda says, is like concentrating on only one section of an orchestra. But with larger fields of view, "a good chunk of the cortical surface of a mouse" becomes open to scrutiny. Neurons flicker like grainy greyscale fireflies, and by tracing those flashes over time, researchers can identify correlations between cells, circuits and larger brain areas. "We can now interrogate what we refer to as multi-regional circuits or multi-regional interactions in real time."

The 2p-RAM system has been licensed to Thorlabs, a microscopy vendor in Sterling, Virginia. One of the first commercially available instruments was bought by neuroscientist Mackenzie Mathis, at Harvard University in Cambridge, Massachusetts. At the Society for Neuroscience annual conference last month in Chicago, Illinois, Mathis presented data to a Thorlabs user-group meeting showing that she could use that system, plus some home-built deep-learning software, to study mice interacting with a video game. "Joystick pulls can be accurately decoded from neural activity," she told the audience.

The RAM in 2p-RAM stands for 'random-access mesoscopy': the system can rapidly move the laser around in the overall field of view. "What that means in practice is that I can, say, go to the motor cortex and record layer-5 output neurons, and then go to another area of cortex and simultaneously image layer 2/3," Mathis says. Such data can reveal how brain regions interact and communicate as the mice play the game.

But the 2p-RAM cannot scan those regions literally simultaneously; there is a delay of several milliseconds as the laser hops from place to place. Helmchen's and Smith's designs use beam-splitters to provide effectively simultaneous imaging of multiple points – a process called temporal multiplexing. Smith and neuroscientist Jerry Chen, who was the lead author of Helmchen's paper and is now at Boston University in Massachusetts, are collaborating on a second-generation system that they say will be able to access four regions within a 2p-RAM-sized field of view simultaneously. Thorlabs, which has already installed more than 25 mesoscopes in various labs, is developing an updated system capable of simultaneous imaging at multiple depths, says Sam Rubin, the company's general manager.

The oblique approach

At Columbia University in New York City, biomedical engineer Elizabeth Hillman has worked out a way to apply light-sheet microscopy to opaque samples. In conventional light-sheet microscopes, the laser beam is focused by its own objective lens, which is perpendicular to the detection objective. This

arrangement limits the size of the sample that the system can accommodate, and precludes its application to live mice.

Hillman's design, called SCAPE 2.0, eliminates one of the objectives, projecting a plane of light obliquely into the sample and capturing the resulting fluorescence with the same lens. The only moving element is a steering mirror, and the system can record volumes at blazing speeds when used in tandem with a fast camera. "We can do three dimensions faster than [point-scanning microscopes] can do two," Hillman says⁷.

But light sheets still cannot penetrate opaque samples well. So Hillman is now developing a two-photon variant that will be able to probe hundreds of micrometres into opaque samples, as well as a mesoscale version for larger samples.

The system can also be applied to another rapidly growing area of microscopy: the imaging of large, 'cleared' tissues, such as mouse brains, that have been made transparent through chemical treatment. Hillman's team was able to image a piece of mouse brain measuring $8.4 \times 9.1 \times 0.4 \text{ mm}^3$ in just 4 minutes⁷. Other light-sheet designs can also tackle such samples. One, called the mesoSPIM, can image a 21-mm field of view⁸; another, developed by physicist Reto Fiolka at the University of Texas Southwestern Medical Center in Dallas, and his colleagues, tiles millimetre-sized fields to capture centimetre-sized samples with subcellular resolution⁹. Neuroscientist Raju Tomer, a Keller lab alumnus and colleague of Hillman's at Columbia, has developed yet another geometry, called 'light-sheet theta microscopy', in which the excitation objective is positioned at an oblique angle to remove light-sheet microscopy's lateral constraint¹⁰. This design provides a 1-mm field of view, but can accommodate samples of theoretically any width.

As these and other designs percolate through the microscopy community, new research avenues will open. But, Keller warns, to have a truly broad impact, such developments will need to be paired with better sample preparation and handling, ease of use and affordability. "If the goal is to image a large sample at high resolution, the microscope can only do so much," he says.

Jeffrey M. Perkel is technology editor at *Nature*.

1. McConnell, G. et al. *eLife* **5**, e18659 (2016).
2. Fan, J. et al. *Nature Photonics* **13**, 809–816 (2019).
3. McDole, K. et al. *Cell* **175**, 859–876 (2018).
4. Chen, J. L., Voigt, F. F., Javadzadeh, M., Krueppel, R. & Helmchen, F. *eLife* **5**, e14679 (2016).
5. Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. *eLife* **5**, e14472 (2016).
6. Stirman, J. N., Smith, I. T., Kudenov, M. K. & Smith, S. L. *Nature Biotechnol.* **34**, 857–862 (2016).
7. Voleti, V. et al. *Nature Methods* **16**, 1054–1062 (2019).
8. Voigt, F. F. et al. *Nature Methods* **16**, 1105–1108 (2019).
9. Chakraborty, T. et al. *Nature Methods* **16**, 1109–1113 (2019).
10. Migliori, B. et al. *BMC Biology* <https://doi.org/10.1186/s12915-018-0521-8> (2018).



Where I work Jean-Pierre Bourguignon

Photographed for *Nature* by
Mashid Mohadjerin

The only painting in my office is a copy of a piece of Chinese calligraphy made by the late Shiing-Shen Chern, one of the most important mathematicians of the twentieth century, which I received from colleagues in 2007 as a birthday gift. It says “Mathematics is fun.”

When I became president of the European Research Council (ERC) in 2014, I thought this piece would inspire me in my role, so I displayed it in my office above a book featuring Chern’s portrait. As president, I maintain regular contact with the European Parliament and the Council of the European Union, which is made up of ministers from the various member states. But for me, it is extremely important to spend at least half my time talking to scientists to understand their needs. I also chair the ERC Scientific Council, which defines the ERC’s funding strategy and promotes scientific creativity and innovation.

Chern approached science as a collaborative, creative endeavour. I met him many times, and his interest in my early mathematical research helped me gain

confidence in my work to understand the role of curvature in geometry. In my career, I have benefited so much from interactions with people from many different countries, so I feel like I have to continue this chain of creating opportunities in the EU for foreign researchers to collaborate. While I have been president, the ERC has signed ten cooperation agreements with countries outside Europe – including China, Japan, Singapore and India – to fund their researchers who come to the EU.

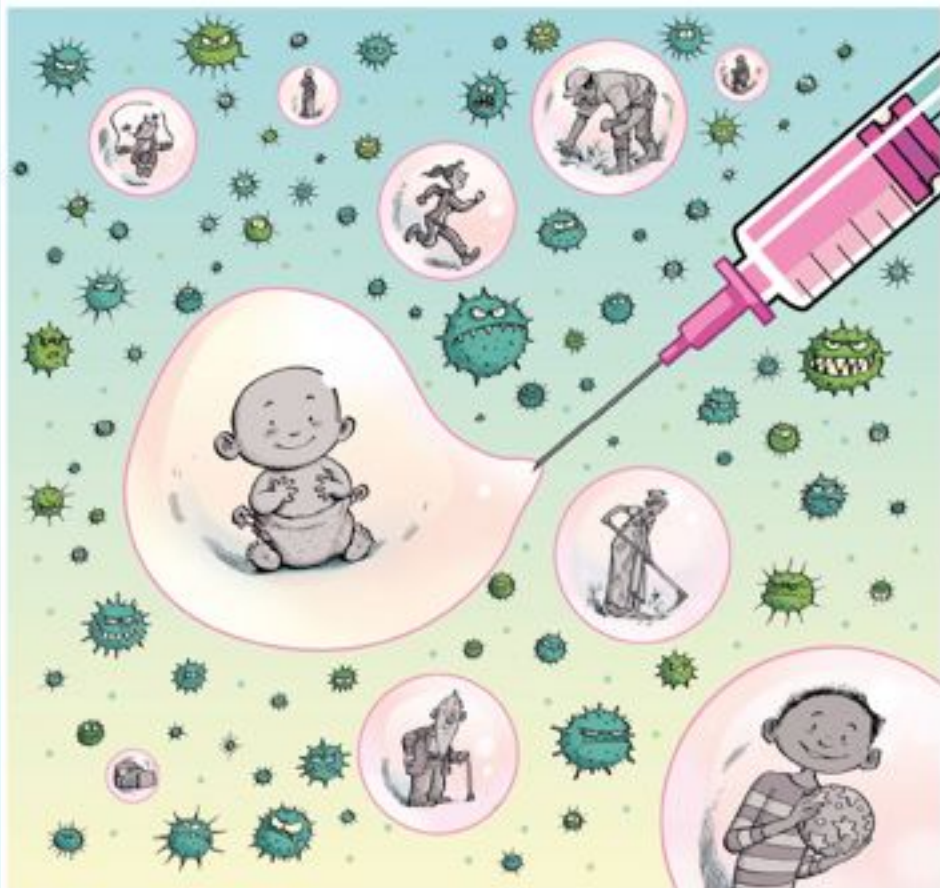
In my office, I have a few plants, a table for holding small meetings, an EU flag and a map of the world. Above the map is a depiction of one of the mottoes of the ERC: “Open to the world.” At the moment, it is unclear how relations with the United Kingdom will proceed post-Brexit. Still, with scientists from more than 80 nations holding ERC grants, we aim to continue making that motto a reality.

Jean-Pierre Bourguignon is president of the European Research Council in Brussels.
Interview by Virginia Gewin.

nature

outlook

Vaccines



Better protection
for everyone

Produced with support from:



Vaccines



For more on immunizations visit nature.com/collections/vaccines-outlook

Editorial

Herb Brody, Richard Hodson, Jenny Rooke, Anne Haggart

Art & Design

Mohamed Ashour, Denis Mallet, Kate Duncan

Production

Nick Bruni, Karl Smart, Ian Pope, Kay Lewis

Sponsorship

Stephen Brown, Nada Nabil, Claudia Danci

Marketing

Nicole Jackson

Project Manager

Rebecca Jones

Creative Director

Wojtek Urbanek

Publisher

Richard Hughes

VP, Editorial

Stephen Pincock

Managing Editor

David Payne

Magazine Editor

Helen Pearson

Editor-in-Chief

Magdalena Skipper

In terms of the public-health benefits that vaccination has delivered, it is almost without an equal – only the provision of safe drinking water has had a greater impact. The World Health Organization estimates that vaccines prevent between 2 million and 3 million deaths from infectious diseases every year. The protection afforded by vaccination is clear, but so is what happens when vaccine coverage in a population falls below the level required to achieve ‘herd immunity’ (see page S44). There are also numerous infections without a vaccine and these continue to claim lives. But researchers are making significant steps towards filling these protective gaps (S46).

Diseases caused by parasitic infection have proved a particularly difficult nut to crack. After decades of research, a vaccine for malaria is being piloted in children in Africa. Although this is a hopeful development, it does not end the quest. The vaccine is imperfect, and other types are being pursued (S51). Progress is also being made in protecting people from parasitic worms (S54). And researchers are exploring the possibility of harnessing the ability of some plants and insects to pass immunity to their offspring to protect them from infection by parasites and other organisms (S55).

In humans, newborns and older people have most to gain from vaccination, because they are the most vulnerable to infectious disease. Unfortunately, vaccines tend to be least effective in these groups. A better understanding of immunity in the old and the very young could lead to vaccines tailored to their needs (S48). Such advances will only bear fruit, however, if people take up the option of vaccination. Groups opposed to the practice have existed for almost as long as the vaccines themselves. For many years, governments have proposed penalizing those who disregard their recommendations, and such mandates are now widespread; evidence of their effectiveness, however, is unclear (S58). For many, a better use of time and money is to listen to the concerns of the hesitant – a much larger group of people than those who are vehemently opposed (S57).

We are pleased to acknowledge the financial support of GlaxoSmithKline plc in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Richard Hodson
Supplements editor

Contents**S44 VACCINES**

Arming the immune system
Vaccines explained

S46 CLINICAL TRIALS

Research round-up
Latest results from vaccine trials

S48 AGE

Vaccinating the vulnerable
Older people and the very young are underserved groups

S51 MALARIA

The problematic parasite
Researchers are struggling to agree on the best approach

S54 Q&A

Taking on worms
Jeffrey Bethony explains why a vaccine for worms is needed

S55 INVERTEBRATES

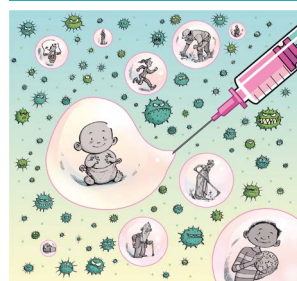
Generation game
What insects can teach us about transgenerational immunity

S57 Q&A

A need for nuance
Heidi Larson on the importance of engaging with parents

S58 MANDATORY VACCINATION

Forcing the issue
Why not everyone is convinced by the idea

**On the cover**

Vaccines offer protection against disease throughout our lives. Credit: David Parkinson

About Nature Outlooks

Nature Outlooks are supplements to *Nature* supported by external funding. They aim to stimulate interest and debate around a subject of particularly strong current interest to the scientific community, in a form that is also accessible to policymakers and the broader public. *Nature* has sole responsibility for all editorial content – sponsoring organizations are consulted on the topic of the supplement, but have no influence on reporting thereafter (see go.nature.com/2NqAZ1d). All *Nature Outlook* supplements are

available free online at go.nature.com/outlook

How to cite our supplements

Articles should be cited as part of a supplement to *Nature*. For example: *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2019).

Contact us

feedback@nature.com
For information about supporting a future *Nature Outlook* supplement, visit go.nature.com/partner

Copyright © 2019 Springer Nature Ltd. All rights reserved.

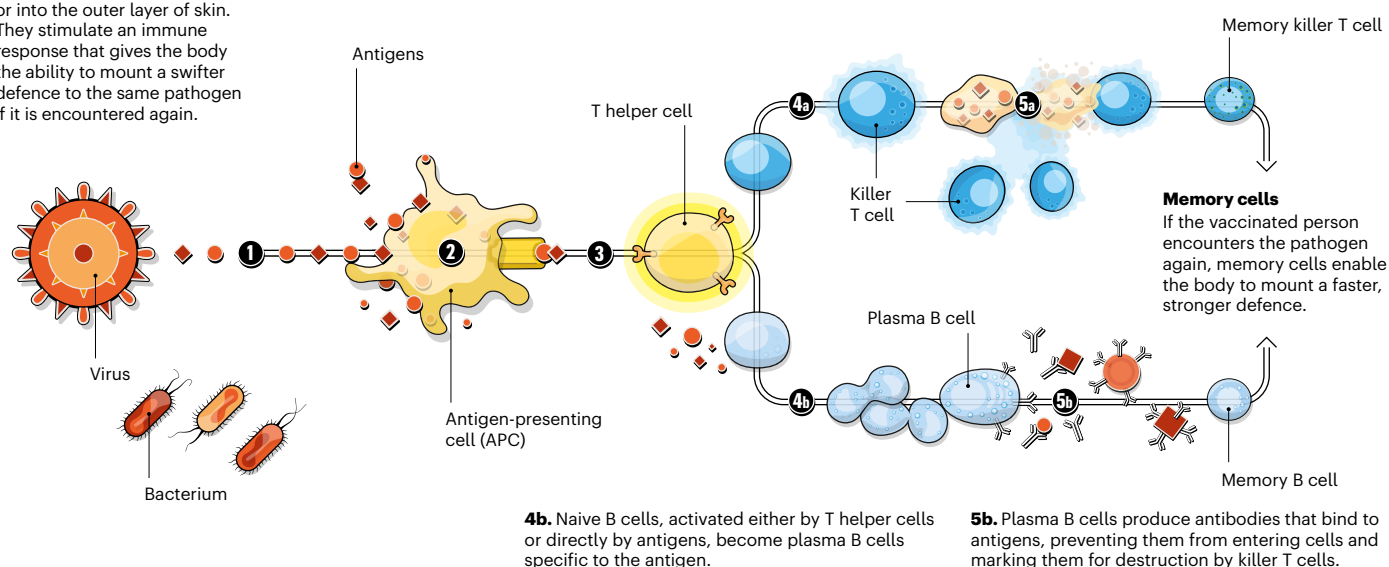
ARMING THE IMMUNE SYSTEM

In 1796, English physician Edward Jenner introduced the first vaccine, for smallpox, when he infected a young boy with cowpox. In the years since, vaccines – a name derived from the Latin word for cow – have been developed for many diseases, saving millions of lives. But the fight to conquer infectious disease continues.

By Neil Savage; infographic by Alisdair Macdonald

IMMUNE STIMULATION

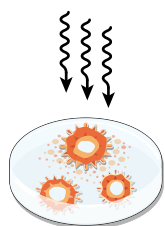
Vaccines can be administered orally, nasally or by injection: into the muscle mass, into the layer between skin and muscle or into the outer layer of skin. They stimulate an immune response that gives the body the ability to mount a swifter defence to the same pathogen if it is encountered again.



VACCINE VARIETIES

There are several types of vaccine in use, each with their own strengths and weaknesses. Many are administered along with adjuvants – substances such as aluminium salts, lipids and RNA that strengthen the immune response.

Inactivated vaccine

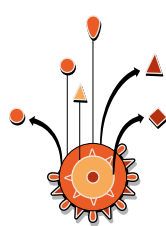


The pathogen is treated with heat or chemicals to kill it before it is introduced into the body.

- ✔ Easy to store and transport.
- ✔ Low risk of causing an infection.
- ✖ Elicits weaker immune response.
- ✖ May require several doses and boosters.

Examples: polio, hepatitis A, rabies

Subunit vaccine

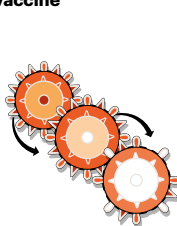


One or more parts of the pathogen, such as a protein, are isolated and used to evoke an immune response.

- ✔ Low risk of adverse reaction.
- ✔ Can be used in people with weakened immune systems.
- ✖ Can be difficult to manufacture.
- ✖ May require boosters.

Examples: hepatitis B, influenza, pertussis

Live, attenuated vaccine

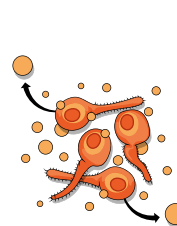


A virus is weakened, often by repeatedly passing it through a tissue culture in which it replicates poorly.

- ✔ Activates killer T cells.
- ✔ One or two doses can provide lifelong immunity.
- ✖ Must be refrigerated.
- ✖ Less safe for people with weakened immune systems.

Examples: measles, mumps, rubella, varicella, rotavirus

Toxoid vaccine



A toxin produced by the pathogen, instead of the pathogen itself, is deactivated and used to produce the immune response.

- ✔ Unable to cause disease or to spread.
- ✔ Stable, so easy to distribute.
- ✖ May require boosters to maintain immunity.

Examples: diphtheria, tetanus

IN DEVELOPMENT

DNA vaccine

DNA from pathogens, sometimes attached to another virus or bacterium, is used to generate an immune response.

In human trials for herpesvirus, influenza and Zika virus.

Recombinant vector vaccine

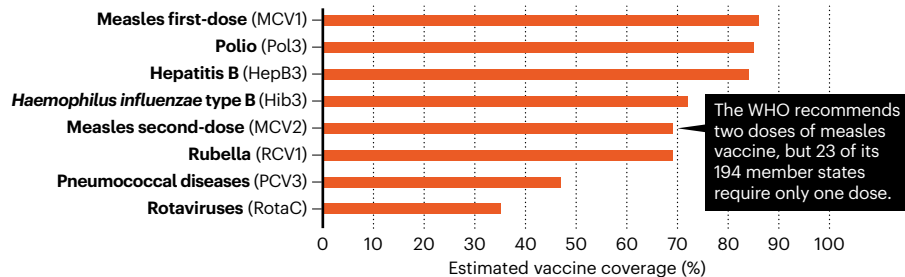
Live but harmless viruses are genetically engineered to express an antigen to a dangerous virus, which the immune system can target.

Being explored for Zika virus, HIV and Ebola.

BUILDING COVERAGE

Vaccines initially developed in the 1950s and 1960s, such as those for polio and measles, are commonly administered globally. Those introduced more recently, such as that for rotavirus, are less widely used.

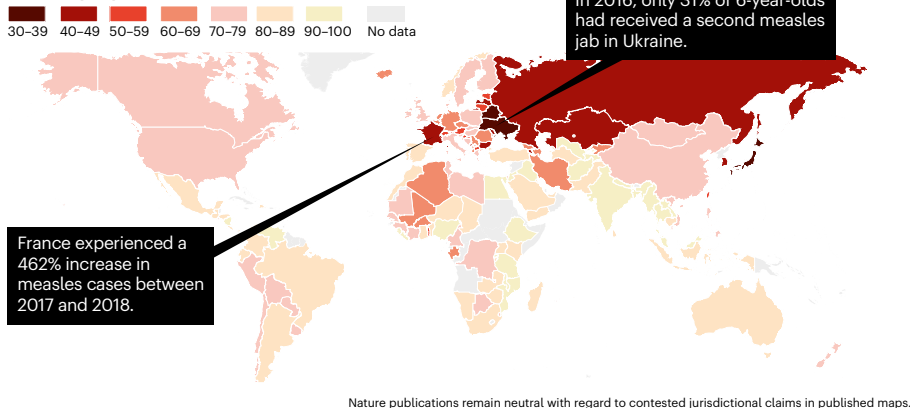
Global vaccine coverage in 2018



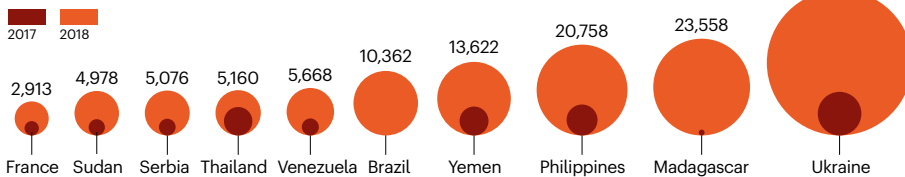
THE LUXURY OF HESITANCY

Although 79% of people globally think that vaccines are safe, trust varies widely between nations. Europe has some of the lowest levels of perceived safety — a finding that might partly explain the surge in measles cases seen in Ukraine in 2018. But other factors besides hesitancy to vaccinate also affect the spread of infections.

Share of people who think vaccines are safe in 2018 (%)

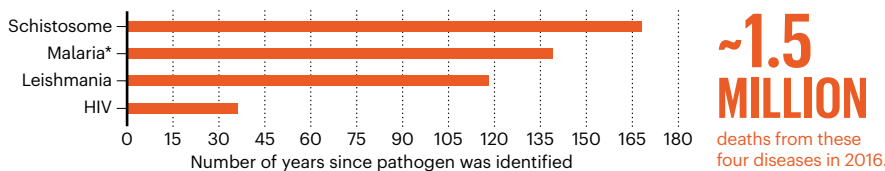


Increase in measles cases 2017-18



LONG ROAD TO NEW VACCINES

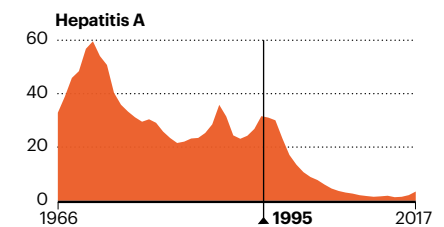
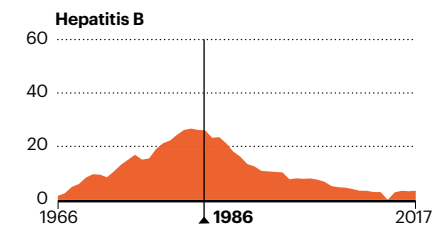
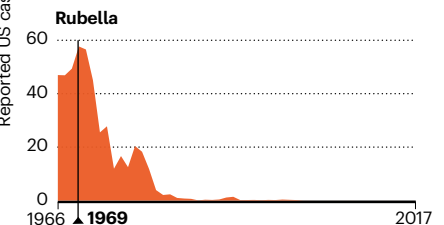
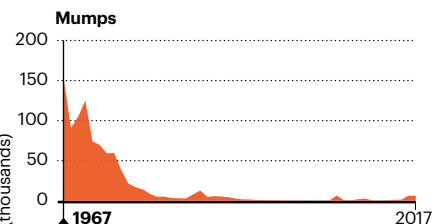
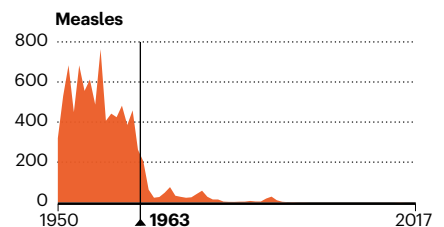
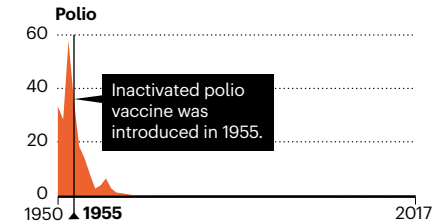
Development programmes are under way for a number of deadly diseases that currently lack a vaccine. In many cases, the pathogen responsible was identified decades ago, but effective vaccination strategies have proved elusive.



*A pilot study of a malaria vaccine began in three countries in 2019.

SAVING LIVES

Case numbers of certain infectious diseases in the United States dropped precipitously after effective vaccines for each were widely adopted.



SOURCES: Immune stimulation: WHO/CDC/www.historyofvaccines.org; Vaccine varieties: US Dept. Health & Human Services/www.historyofvaccines.org; Building coverage: WHO; The luxury of hesitancy: Gallup Wellcome Global Monitor 2018 (Wellcome, 2019); Long road to new vaccines: S. Vanderslott & M. Roser <https://ourworldindata.org/vaccination> (2019)/WHO; Saving lives: S. W. Roush et al. *J. Am. Med. Assoc.* **298**, 2155–2163 (2007)/CDC.

Research round-up

Highlights from vaccine trials. By Elizabeth Svoboda

Mosaic approach to HIV vaccine trials

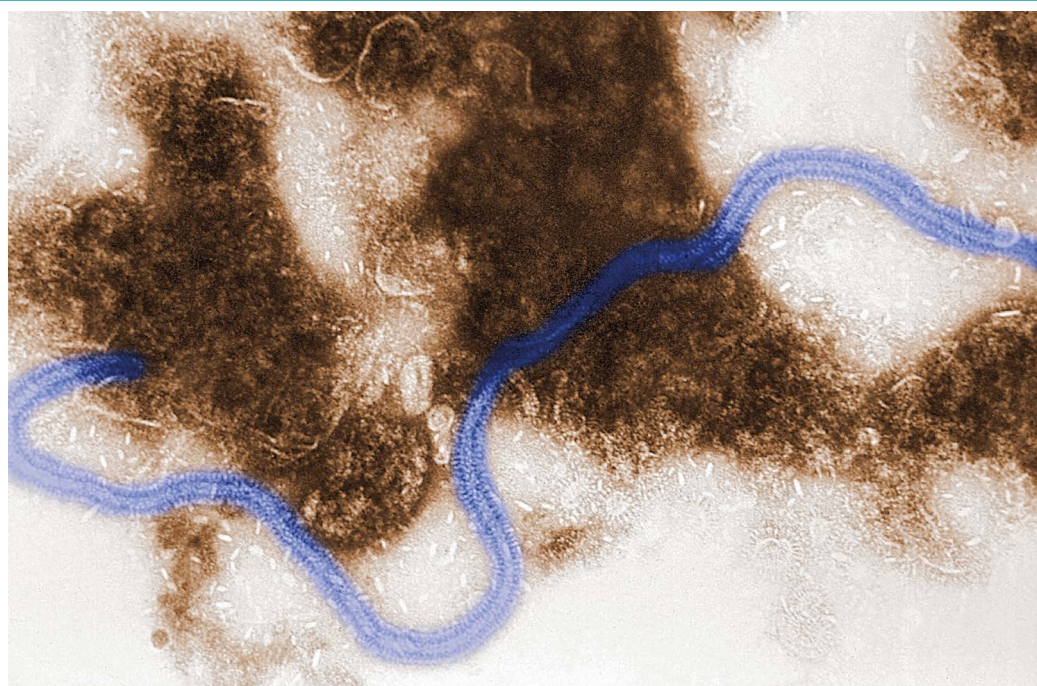
A vaccine for HIV has long been elusive because the virus comes in many different guises. A 'mosaic' vaccine, which includes snippets of variants of HIV from around the world, now entering a phase III trial could change that.

In a phase I/II trial published last year and led by Dan Barouch at Harvard Medical School in Boston, Massachusetts, more than 80% of participants showed a robust immune response to the best-performing version of the experimental vaccine. The volunteers produced a range of antibodies that bound specifically to different HIV strains, and the researchers saw clear evidence of phagocytosis, in which immune cells surround and digest cells infected with the HIV virus.

These outcomes have raised hopes that this vaccine – set to be the fifth tested for efficacy in humans – will perform well in a phase III trial. The trial, scheduled to start before the end of the year, is aiming to recruit 3,800 men and transgender people who have sex with men, transgender people or both. Unlike the phase I/II trial, this trial will show whether the vaccine prevents people from getting the virus. Results are expected in 2023.

Although antiretroviral therapy has reduced the burden of HIV around the world, the stakes for these trials remain high, because a safe and reliable vaccine is needed to end the pandemic entirely, say HIV specialists.

Lancet **392**, 232–243 (2018)



Transmission electron micrograph image of human respiratory syncytial virus (blue).

A precision-engineered RSV jab

It might not be as notorious as the flu, but respiratory syncytial virus (RSV), which causes cold-like symptoms, kills 160,000 people globally per year, and is especially serious for infants and older people. At the moment, there is no effective vaccine. However, a new candidate based on the virus, but with a precise tweak to the structure of one of its surface proteins, has shown signs of generating a substantial immune response in early trials in people.

To mount an effective defence against RSV, the body needs to churn out large numbers of antibodies against the virus. Many of those target a protein on its surface – known as the fusion glycoprotein, or F protein – that is crucial to it fusing with host cells. But previous RSV vaccine candidates could not create a sufficient immune response because the F protein changes

shape once the virus enters a cell, rendering antibodies against its pre-fusion form less effective.

Because the pre-fusion form of the F protein generates the stronger immune response, Barney Graham at the US National Institutes of Health's Vaccine Research Center, Bethesda, Maryland, and his colleagues engineered a partial version of RSV with an F protein that could not change shape. The researchers hoped that providing a form of the protein locked in this state would allow the body to generate enough antibodies to ward off future infection.

In a phase I trial, 40 adults who were given the vaccine churned out 7–15 times more RSV-fighting antibodies than were present before vaccination – an increase that persisted for months. This is a larger antibody response than is seen in people after natural RSV infection. Subsequent trial phases, however, must prove not only that the vaccine boosts RSV

antibody levels but also that it either prevents the disease or, at least, reduces its severity.

Science **365**, 505–509 (2019)

Malarial parasite trapped in the blood

An experimental vaccine for malaria can produce antibodies in humans that lock the disease-causing parasite *Plasmodium falciparum* outside red blood cells. Malaria symptoms are the result of the parasite multiplying inside these cells and causing them to burst, so researchers hope that this approach will lessen the damage caused.

Despite an array of drugs and mosquito-killing agents, malaria remains a deadly scourge in low- and middle-income countries. Attempts to create a vaccine against the parasite during the blood stage of its life cycle have flopped because the antibodies

created fail to stop the parasite entering red blood cells. Matthew Higgins at the University of Oxford, UK and his colleagues, therefore, opted to target a crucial parasite protein called PfrH5. This protein binds to the host's red blood cells to allow the parasite access – like a key opening a door. Blocking PfrH5 locks the parasite out of the cells, and thereby prevents it from damaging them.

During a phase I trial, the scientists collected the blood of people who'd received the vaccine, sequenced the genes that produced their antibodies, and used the genes to make more antibodies. The researchers then tested how strongly the antibodies reacted to the PfrH5 protein. Of the 17 distinct antibodies they found, 7 strongly inhibited the parasite by preventing PfrH5 from binding to red blood cells. The team also identified another antibody that slows PfrH5's binding rate. Although this antibody doesn't stop PfrH5 from binding, it is still helpful because it buys time for the other antibodies to act.

The team is now planning a version of the vaccine that creates more of the antibodies that block or stall PfrH5, and fewer of the ones that have little or no effect.

Cell **178**, 216–228.e21 (2019)

Antibiotics lower vaccine effectiveness

Antibiotics wreak havoc on the microbes in our gut. Signature side effects include diarrhoea or constipation, but the toll might be more than just digestive. A study published in September suggests that the depletion of gut microbes that follows antibiotic use can make the influenza vaccine less effective in people with low natural immunity.

During a two-part trial, Bali Pulendran at Stanford University in California and his colleagues

gave flu vaccines to a total of 33 adults. Participants in the second phase of the trial hadn't encountered the virus recently or received the flu vaccine in three years, and were therefore considered to have low immunity. The scientists gave half the participants a course of broad-spectrum antibiotics, including neomycin, vancomycin and metronidazole. Gut-microbe

“People who take a course of antibiotics might have a weaker immune boost from the flu injection.”

diversity plummeted in all volunteers treated with antibiotics, but people with low levels of flu immunity also produced very few antibodies in response to the flu vaccine, meaning they might be more prone to develop the disease if exposed to the virus. People who did not receive antibiotics in either phase, however, displayed a normal antibody response to the flu jab, showing they were protected from the vaccine strains.

The results hint that some people who take a course of antibiotics might have a weaker immune boost from their yearly flu injection. Although the mechanism is not yet clear, the researchers note that production of bile acids such as lithocholic acid dropped 1,000-fold in people treated with antibiotics. Normally, gut bacteria help to manufacture these acids, which are known to regulate immune activity. The researchers think that when gut microbes are depleted, impaired lithocholic acid production might interfere with the body's ability to create a normal immune response to the flu vaccine.

Cell **178**, 1313–1328 (2019)

A bulwark against chlamydia

More common than gonorrhoea, syphilis and HIV combined, chlamydia has been the subject of vaccine research for more than half a century – without much success. But now a candidate vaccine for the sexually transmitted bacterial infection, which can cause infertility and chronic pain, has generated a strong immune response in early trials in people.

Previous uses of weakened chlamydia bacteria have failed to produce enough antibodies for long-lasting immunity. So Peter Andersen at Statens Serum Institut in Copenhagen and his colleagues isolated a protein on the surface of the bacterium called the major outer membrane protein, which had evoked a strong antibody response in animal tests. The team then tweaked its structure so that it could generate immunity to multiple strains of the bacterium.

In a phase I trial, the researchers gave 35 women aged 19 to 45 either a version of the vaccine that included aluminium hydroxide, or one that included lipid molecules or a placebo. Each participant received five doses of vaccine – three injected and two sprayed into the nose. Although both the aluminium hydroxide and the lipid-molecule variants created a robust immune response in all the volunteers compared with the placebo, the vaccine with lipid molecules performed best, generating more than five times as many chlamydia antibodies as did the other formulation.

The researchers are planning a phase II study of the lipid-variant vaccine. If its effectiveness is confirmed in further trials, it will be of particular benefit to girls and young women in low- and middle-income countries – a demographic with a high incidence of chlamydia. However, because the vaccine is based on an engineered protein rather than a live attenuated

version of the bacterium, it could prove expensive to produce.

Lancet Infect. Dis. **19**, 1091–1100 (2019)

CAR-T takes on solid tumours

Lab-modified immune cells called chimeric antigen receptor (CAR) T cells are widely used to treat blood cancers. Darrell Irvine at the Massachusetts Institute of Technology, Cambridge, and his colleagues have developed a vaccine that could allow CAR-T cells to also attack solid-tumours.

Solid tumours tend to suppress immune cells. To address this problem, the team joined a specific antigen that boosts CAR-T cell activity to a molecule with a lipid tail that hooks onto albumin proteins in blood. Once in the blood, the vaccine molecules attach themselves to albumin and are carried to the lymph nodes, which regulate the body's immune responses. The molecules' lipid tails penetrate the surfaces of lymph-node cells, and the embedded surface antigens stimulate the activity of tumour-fighting CAR-T cells.

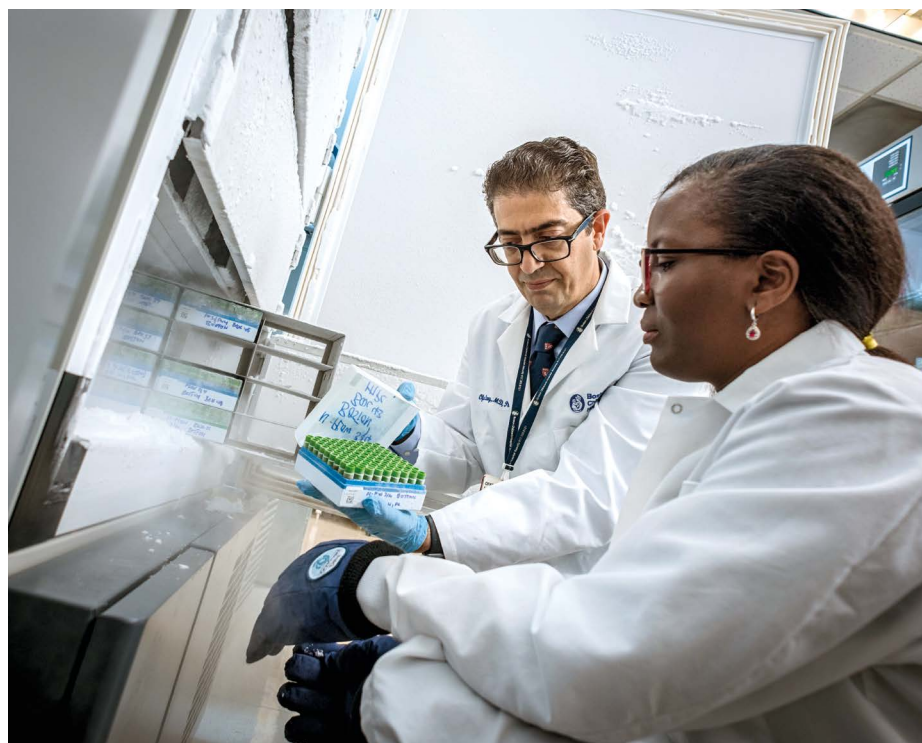
The team tested the vaccine in mice with glioblastoma, melanoma and breast tumours. In about 60% of mice that received the vaccine with an infusion of CAR-T cells, tumours disappeared completely. Tumours did not shrink in mice that received only CAR-T cells.

The authors also found that human cells studded with CARs rev up activity of tumour-fighting cells, indicating the vaccine could prove viable in people. The developers of the approach plan to test it in clinical trials in the next few years.

Science **365**, 162–168 (2019)



For the latest research highlights published by *Nature* visit [nature.com/research-highlights](https://www.nature.com/research-highlights)



Ofer Levy and his colleague look at biosamples collected from infants in Gambia.

Vaccinating the vulnerable

Researchers are tailoring immunization strategies to improve vaccine responses among newborns and older people. **By Amanda Keener**

In the United States, 70% of people who are hospitalized for influenza are 65 or older. Older people are commonly encouraged to have a flu vaccine each year. But, in what seems like a cruel twist of fate, vaccines against the flu and other respiratory diseases are less effective in older people. Depending on the year, the flu jab typically protects 30–60% of middle-aged recipients; for those aged 65 and over, the protection rate is more like 20–50%.

Infants, particularly newborns, are at a similar disadvantage. Their symptoms are typically worse than those of older children and adults. But again, most vaccines offer less protection to newborns than to adults, and few are recommended before 8 weeks of age. For instance, the vaccine against the bacterium *Haemophilus influenzae*, which can

cause meningitis and sepsis, generates a poor immune response in babies under 2 months.

It isn't that those at the far ends of the age spectrum don't respond to vaccines at all, but that they respond differently – and most vaccine testing is done on young and middle-aged adults. “Vaccines have traditionally been developed as one-size-fits-all, but if the immune response is different when you're young and old, that approach leads to lots of failure,” says Ofer Levy, who directs the Precisions Vaccine Program, which looks at how variables such as age affect vaccine responses, at Boston Children's Hospital in Massachusetts. This way of developing vaccines is a disservice to the people most in need of protection, says Tobias Kollmann, a vaccinologist at Telethon Kids Institute in Nedlands, Australia. “If you look at the current

vaccine schedule, who gets most of the vaccines? The very young and the very old. And yet we don't understand a thing about them.”

To better protect vulnerable populations, researchers are investigating how the immune system changes with age. “We want to learn the molecular rules of why vaccines may or may not work in different age groups,” Levy says. He and Kollmann are building a biological signature of what a strong vaccine response looks like in newborns. Others are trying to sidestep the unique properties of the newborn immune system by protecting them before birth, through their mothers (see ‘Maternal instincts’). And some research groups and companies are tailoring vaccines to boost the responses of older people. Collectively, the aim is to design vaccines and immunization strategies that work for the people most in need.

Early protection

The freezers in Levy's lab are filled with thousands of samples of serum, the liquid component of blood after cells and clotting factors are removed. Most of the samples were taken in the first week of a child's life – some from infants born across the street at Boston Children's Hospital, others from as far away as Papua New Guinea. Together, they hold clues about the newborn immune system that Levy and his collaborators hope will lead to sorely needed improvements in vaccine design.

According to the World Health Organization (WHO), worldwide, 47% of all deaths before the age of five occur during the first four weeks of life, and infections are responsible for 20% of those deaths. Effective vaccines could see more infants through that difficult first month, says Beate Kampmann, an infectious-disease specialist at the London School of Hygiene and Tropical Medicine who directs vaccine research at the UK Medical Research Council unit in Gambia. “There's a myth out there that the baby's immune system doesn't respond to vaccination, and that's not true,” she says. Newborns respond well to the bacillus Calmette–Guérin (BCG) vaccine against tuberculosis, the oral polio vaccine and hepatitis B vaccines. The challenge is to work out why those vaccines work, and use that information to create or retool others.

Both hepatitis B and BCG vaccines circumvent the newborn immune system's relative tolerance of foreign organisms. At birth, babies leave the essentially sterile environment of the womb and are bombarded by microorganisms, most of which are beneficial (or at least benign). Because treating all of these organisms as invaders would use so much energy that there would be little left for growth, newborns' immune responses are

SAM OGDEN

blunted. Their innate immune cells – the ‘first responders’ that detect common patterns on bacteria and viruses – make lower levels of antiviral and antibacterial molecules compared with immune cells in adults. This lack of response is a problem for vaccines, because, in general, more robust activation of the immune system means it remembers the antigens it was exposed to for longer.

Newborn immunity is characterized by rapid flux. Even in the first few days of life, major changes occur in genes and cells related to immunity that researchers are just beginning to identify. Kollmann says that getting a handle on how best to immunize newborns requires a detailed map of those changes. That’s where Levy’s freezers come in. He and other members of the Expanded Program on Immunization Consortium (EPIC) – an academic group that focuses on vaccination studies in infants – are cataloguing every gene, cell and protein regulated by the immune system that they can find in blood collected during the first week of life. Their study includes samples collected by Kampmann’s team in Gambia and by collaborators at another field site in Papua New Guinea.

In a study published earlier this year involving 30 newborn babies, the EPIC team documented marked changes in the quantities of several types of immune cell during the first week of life¹. These data allowed the researchers to map the common developmental trajectory of the babies’ immune systems over that week. Kollmann likens the result to the charts that health professionals use to monitor childhood growth, and suggests this approach could eventually be used to test how potential interventions, such as supporting women so that they can breastfeed for longer, or giving infants probiotics, affect immune health and vaccine responses. The team is adding several hundred more samples to the analysis, so that it can see how an infant’s vaccination history affects immune-system development.

Superior shots

The consortium’s other major goal is to understand why some infants respond better to vaccines than others. “Now we’ve got the tools to dissect why things work and why they don’t work,” Kampmann says. “That is really going to reform the way that we think about the next generation of vaccines.” The routine hepatitis B vaccine, for example, is typically given in three parts during a baby’s first year or so, but around one-third of infants are fully protected from the virus after just one injection, says Kollmann. Achieving this ‘one dose protection’ for more vaccines would be a huge win for public health in much of the world, he

Maternal instincts

The importance and difficulty of protecting infants in the earliest days of life has led to the development of an alternative vaccination strategy for some diseases. The idea is to make use of a natural phenomenon in which antibodies – proteins that help the body to recognize and attack invaders such as viruses and bacteria – are transported across the placenta from mother to developing fetus. After birth, those antibodies act “like a shield that is around the baby for the first three months or so” says Beate Kampmann, an infectious-disease specialist at the London School of Hygiene and Tropical Medicine. Maternal vaccination stimulates the production of antibodies in the mother that are passed to newborns and can protect them from the severe symptoms of infections such as whooping cough and flu before they can receive their first vaccinations at two months.

Questions remain, however, about how best to implement this form of protection. For example, it is unclear how the timing of vaccination or a mother’s HIV status affect the number of antibodies that cross the placenta. There is also some concern that maternal vaccination might depress an infant’s responsiveness to the same vaccine later on – maternal antibodies might mask the vaccine and prevent the child’s immune cells from recognizing it. To clarify this, Kampmann is leading a study in Gambia that

will measure responses to the whooping cough vaccine in 600 children whose mothers were vaccinated against the disease during pregnancy.

Kampmann directs the Immunising Pregnant Women and Infants network (IMPRINT), which supports research to advance maternal vaccines. Two infections for which such vaccines are making strides are respiratory syncytial virus (RSV), a common infection that can be serious for infants with other health problems, and group B streptococcus, which babies pick up from their mothers at birth and that caused 90,000 deaths globally in 2015 (ref. 8). Several group B streptococcus vaccine candidates are now being tested, including one that has been tested in a phase II trial in pregnant women in Malawi and South Africa. The maternal vaccine induces antibody responses in infants, but it’s not clear yet whether those responses are enough to prevent disease⁹. The most advanced RSV vaccine in development is a maternal vaccine developed by Novavax in Gaithersburg, Maryland. In a phase III trial, the vaccine proved about 40% effective at preventing RSV infection in the first 90 days of life. “That certainly means that we’re on the right track,” says Justin Ortiz, an epidemiologist at the University of Maryland School of Medicine in Baltimore, who was not involved in the study.

says. In many low- and middle-income countries, parents must carry their children for kilometres on foot to reach vaccine stations, he explains. It’s a trek that they might not be willing or able to make multiple times, so fully protecting children with a single dose could save millions of lives every year, he says.

As part of an ongoing study in Gambia, the EPIC team is trying to identify the immune signatures of children who are fully protected from hepatitis B after one injection, so that vaccines can be designed with one-dose protection in mind. To generate these signatures, the team will carry out the same types of analysis done for their immune-health trajectory study. It will examine blood collected before and at several time points after hepatitis B vaccination, and look at changes in the number of immune cells in the blood, cytokine concentrations and gene expression.

The researchers are also testing the use of

another vaccine as an enhancer, or adjuvant, for the hepatitis B vaccine. Researchers have known for decades that infants who receive the BCG vaccine at birth are also less likely to die from a host of other infections. The BCG vaccine is thought to heighten innate immune-cell sensitivity and enhance responses to other vaccines such as pneumococcal and tetanus jabs². Last year, Kollmann, Levy and their collaborators showed that the BCG vaccine boosted the immune response of newborn mice to the hepatitis B vaccine³. The researchers hope that a trial in Gambia, due to be completed in 2021, will show whether and how the BCG vaccine might do the same in infants.

Levy’s team is also on the hunt for synthetic adjuvants that stimulate the infant immune system as powerfully as the BCG vaccine does. This would allow vaccines to be optimally designed for use in newborns, rather than requiring health professionals to combine multiple



A health worker gives the BCG vaccine to a newborn baby in Guinea-Bissau.

vaccines that might not always be available.

In 2017, the team reported that synthetic nanoparticles packed with imidazoquinoline molecules can activate a first-responder protein called Toll-like receptor (TLR) 8, as does the BCG vaccine⁴. The group is now developing a vaccine against the bacterium *Bordetella pertussis* – the cause of whooping cough – with a TLR8-activating adjuvant. It plans to test whether TLR-stimulating adjuvants allow the vaccine to provide lifelong protection from birth – something that the current protein-based vaccine, introduced in the 1990s, does not afford.

Ageing immunity

Babies are unprotected against many vaccine-preventable diseases for months until they can have their first set of vaccines, but those at the opposite end of the age spectrum are left susceptible for many years – especially people with declining lung or heart function or those in assisted living communities, where infections can spread quickly. The outcomes of respiratory infections for people over the age of 65 can be even worse than for infants. It's estimated that in the United States, respiratory syncytial virus kills more than 10,000 people over 65 each year, and hospitalizes three times as many older adults as it does children under five. The flu has an even bigger impact. The US Centers for Disease Control estimates that the 2017–18 flu season caused more than 68,000 deaths among the older people. These numbers are likely to increase, says Gregory Poland, an immunologist at the Mayo Clinic in Rochester, Minnesota. By 2050, the global population of people over 60 is expected to be double what it is today. “Most

people have ignored the silver tsunami that is coming,” Poland says.

Poland says vaccines for older adults will be most successful if they are tailored to the immune characteristics of older people. Along with Kollmann and Levy, he is part of the Human Immunology Project Consortium (HIPC), which is cataloguing the immune signatures of different age groups before and after immunization. In 2017, HIPC researchers reported⁵ that several such signatures that could be used to predict responses to vaccines in adults under 35 could not be used for populations of people over 60. That means that the vaccine formulations that work best for older individuals will be unique to that age group.

As people approach 50, Poland says, detrimental changes to the immune system can already be observed. Beyond 60, he says, “virtually everything that we know to look at becomes compromised”. Some immune cells, he explains, become “exhausted” from chronic activation, including those that keep the varicella zoster virus – the cause of chicken pox and shingles – in check. That's one reason why researchers think the immunity garnered by a childhood bout of chicken pox often fails to prevent shingles past the age of 50.

Several companies are finding ways to overcome the low responsiveness of the ageing immune system. Some are increasing vaccine potency and the number of doses given to elicit a bigger response. A high-dose flu

“Most people have ignored the silver tsunami that is coming.”

vaccine for people aged 65 and over produced by Sanofi Pasteur in Lyon, France, for instance, uses four times more influenza antigen than the standard injection. It is 24% more effective than the regular dose at preventing influenza and influenza-related deaths in those 65 and over⁶. Other vaccines such as a flu vaccine produced by Seqirus in Maidenhead, UK, use adjuvants to kick-start the immune response. The vaccine creates a strong immune response in older recipients, but it's not clear yet whether that translates to better protection⁷.

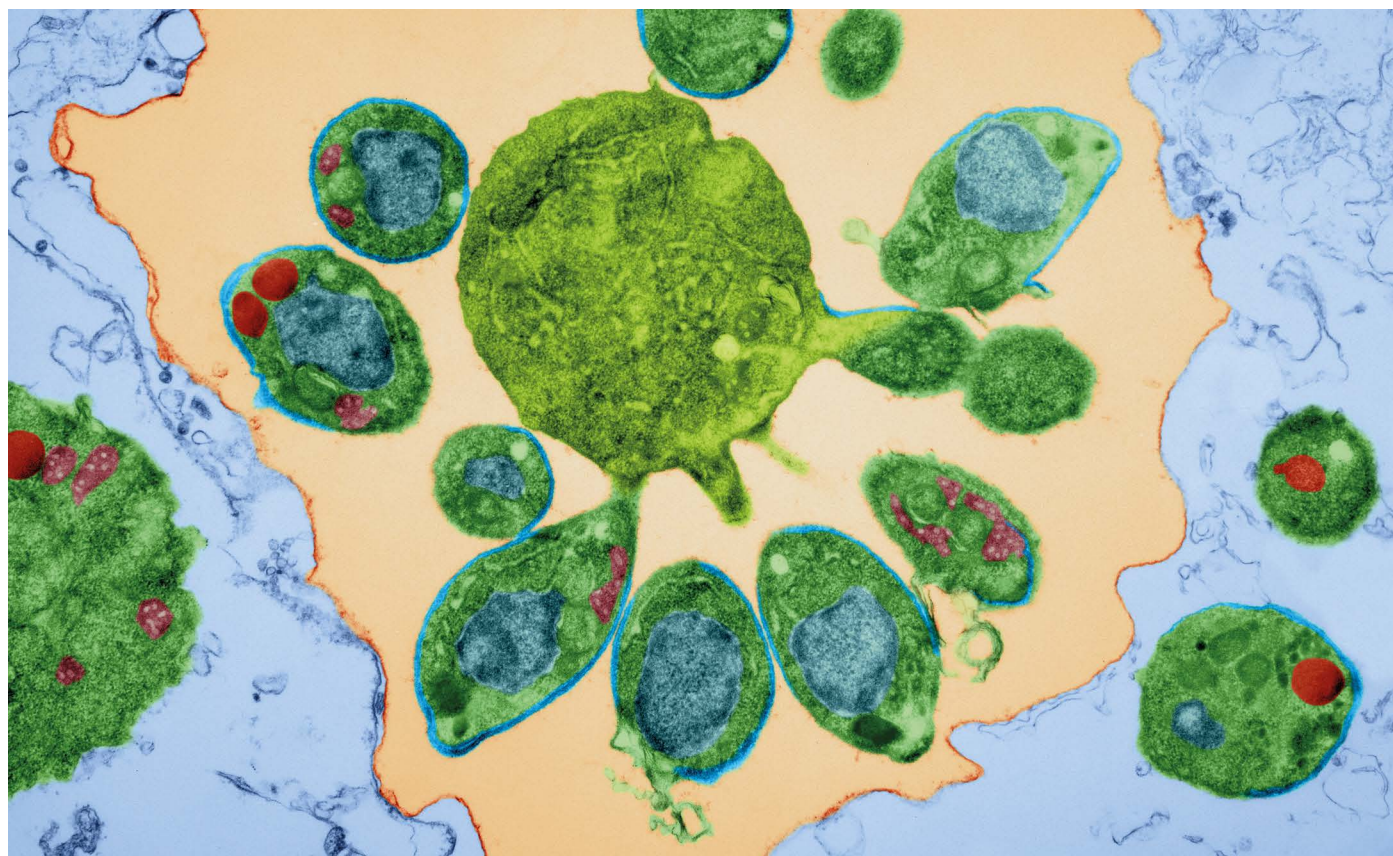
The shingles vaccine Shingrix is a prime example of what's possible when vaccines are built with specific populations in mind. Shingrix contains two adjuvants that stimulate the innate immune system, and is more than 90% effective at preventing shingles in adults aged 50 and over. Just two years since its approval, Shingrix has replaced the previous live-virus vaccine as the preferred shingles vaccine in the United States.

Vaccine design is only part of the effort to protect the old and very young from infection. There's general agreement that vaccines already in use could be saving millions more lives if they were better implemented. Improving flu-vaccine coverage – even if a vaccine performs suboptimally in people over 65 – has the potential to save thousands of lives each year, says Justin Ortiz, an epidemiologist at the University of Maryland School of Medicine in Baltimore. “That can be done now,” he says. Administering existing vaccines to children – the main transmitters of flu – could also reduce infections in older people.

Kampmann agrees that the development of vaccines and immunization strategies for newborns and older people will only be useful if they are paired with improvements in vaccine coverage and access – particularly in low-resource countries, which carry most of the world's infectious-disease burden. “As much as all the biology is really exciting, we should also use our power and advocacy to make sure that the people who need these vaccines get them,” Kampmann says. “We have different fronts to fight.”

Amanda Keener is a freelance science writer in Littleton, Colorado.

1. Lee, A. H. et al. *Nature Commun.* **10**, 1092 (2019).
2. Ritz, N., Mui, M., Balloch, A. & Curtis, N. *Vaccine* **31**, 3098–3103 (2013).
3. Scheid, A. et al. *Front. Immunol.* **9**, 29 (2018).
4. Dowling, D. J. et al. *J. Allergy Clin. Immunol.* **140**, 1339–1350 (2017).
5. HIPC-CHI Signatures Project Team & HIPC-I Consortium *Sci. Immunol.* **2**, eaal4656 (2017).
6. DiazGranados, C. A. et al. *Vaccine* **33**, 4988–4993 (2015).
7. Domnich, A. et al. *Vaccine* **35**, 513–520 (2017).
8. Seale, A. C. et al. *Clin. Infect. Dis.* **65**, S200–S219 (2017).
9. Heyderman, R. S. et al. *Lancet Infect. Dis.* **16**, 546–555 (2016).



DENNIS KUNKEL MICROSCOPY/SPL

Transmission electron micrograph of merozoite-stage malarial parasites, which have caused a red blood cell to rupture.

The problematic parasite

This year, the first vaccine for malaria was given to children. Scientists are working on improvements, but there is little agreement on how to do this. **By Anthony King**

In 1991, when immunologist Patrick Duffy was joining the US National Institute of Allergy and Infectious Diseases, the research community was sceptical that it would be possible to vaccinate against parasites – especially the *Plasmodium* parasites that cause malaria. These organisms, each just a few micrometres across, can shape-shift and hide from the human immune system, moving from the blood into liver cells, before bursting out in a new form to take over red blood cells. They deploy similar cellular machinery to humans, and are much more complex than viral or bacterial foes – for instance, whereas the Ebola virus can encode 7 proteins, *Plasmodium* boasts genes for 5,000.

But in 2019, children in Ghana, Kenya and Malawi began receiving the RTS,S vaccine against malaria as part of a pilot programme.

After 30 years in development, RTS,S is the first malaria vaccine shown to offer protection to young children in a phase III trial – malaria takes 1,200 lives each day, mostly those of children under 5. “The biggest news is that we have malaria vaccines that work,” says Duffy.

RTS,S, developed by pharmaceutical firm GlaxoSmithKline, based in London, works by introducing the immune system to a fragment of a protein that is present on the surface of *Plasmodium* when the parasite enters the bloodstream through an infected mosquito. The protein stimulates the production of antibodies, and allows the body to mount a swift response to the parasite the next time it is encountered. The vaccine’s pilot programme, coordinated by the World Health Organization, is expected to immunize at least 360,000 children a year until 2022. But malaria

researchers are not popping the champagne corks and relaxing just yet. “RTS,S is leading the pack, but it’s a suboptimal vaccine,” says Michael Good, a vaccine researcher at Griffith University in Brisbane, Australia.

“Some feel it doesn’t give you enough protection to be worthwhile,” says Stefan Kappe, an infectious-disease scientist at Seattle Children’s Hospital in Washington. Over the course of the vaccine’s four-year phase III trial, it prevented around 30% of serious cases of malaria¹. The trials also showed a rise in mortality among girls who were vaccinated. “If that is real, that is the end of that story,” says Adrian Hill, a vaccine researcher at the Jenner Institute in Oxford, UK.

Malaria-vaccine researchers are working on several approaches to achieve more robust protection than that offered by RTS,S. Some

are looking to empower the immune system to go after *Plasmodium* in the liver, including using live attenuated parasites as immunizing agents. Others have their eyes on targeting the parasite inside red blood cells – if not to neutralize it entirely, then at least to prevent it spreading to other people. There are numerous strategies all jostling for the limelight, and considerable disagreement about which should be prioritized.

Following the leader

When a person is bitten by a mosquito carrying *Plasmodium falciparum*, the deadliest of the five species of malaria parasite that can infect humans, a handful of protozoa in their needle-like ‘sporozoite’ form enter the body (see ‘Breaking the cycle’). The RTS,S vaccine stimulates the body to produce antibodies against proteins present on the surface of a sporozoite, but the parasite does not linger in this form for long – in half an hour, it can ensconce itself inside liver cells. There, it multiplies, and 7–10 days later emerges as 30,000 merozoite-stage parasites, each invisible to the RTS,S-induced antibodies.

This presents RTS,S with a daunting task.

One sporozoite reaching the liver can result in full-blown malaria, so the vaccine must prime the body with enough antibodies to destroy every single sporozoite before they make it to the liver. It takes a lot, says Hill – around 500 times more antibodies than are produced by a meningitis vaccine. This requires the help of an adjuvant, which chemically stimulates the immune system. RTS,S consists of a region of repeating amino acids from the surface circumsporozoite protein and a hepatitis B surface antigen – a set-up that is the legacy of the vaccine’s long development history. Back in 1987, researchers “couldn’t make what they wanted, which is a virus-like protein with just malaria on the surface”, says Hill. And although technology has improved, once a vaccine has entered clinical trials, you cannot simply overhaul the whole programme.

Hill and his colleagues are working on an alternative vaccine, R21, that has many similarities to RTS,S. R21 fuses together a hepatitis B antigen and half the circumsporozoite protein – a larger portion than in RTS,S². Hill thinks that this combination will be at least as effective as RTS,S, but it is less expensive because it can be administered in

doses one-fifth the size, uses more-modern production methods and has a cheaper adjuvant. A phase II trial in Burkina Faso to test efficacy in adults and children is now under way, with results expected in 2020.

Clear the hideouts

The way in which sporozoite-targeting vaccines prepare the body to mount a defence is only good for a tight window of time after infection. The immune response, therefore, needs to be rapid and effective. An alternative strategy is to target the *Plasmodium* parasite once it is inside liver cells. With this approach, any immune response will have at least five days to eliminate infected liver cells. T cells – immune cells that react to pathogens lurking in cells – can be trained to recognize proteins on the surface of infected liver cells, and kill the cells. The symptoms associated with malaria occur only after the parasite leaves the liver and starts destroying red blood cells, so eliminating infected liver cells would prevent illness and transmission.

Denise Doolan, a vaccine scientist at the Australian Institute of Tropical Health and Medicine in Cairns, has systematically evaluated all of the parasite’s proteins. She looked for proteins that would be spotted by antibodies or T cells found in the blood serum of people with a tolerance to malaria. Now, she has a promising list of proteins associated with protection against malaria that could be useful for vaccine development. “We have three antigens that we have selected from many years of screening that I would love to get to the clinic,” says Doolan. She is especially keen on antigens that elicit a response against both the liver and blood stages of the parasite’s life cycle, and those that show potential to stimulate an immune response in multiple *Plasmodium* species.

Wiping out all *Plasmodium*-infected liver cells requires very large numbers of T cells. One way to ramp up a T-cell response is with a one-two punch: prime the immune system by injecting a circular piece of *Plasmodium* DNA or a viral vector expressing a gene found in *Plasmodium*, and then boost it with adenoviruses engineered to express the same parasite gene. The combination triggers the immune system to muster a robust T-cell response that it will remember in future. But delivering it is logistically challenging because it requires two different vaccines to be administered sequentially. Hill points to encouraging results (67% efficacy)³ from a small trial in Kenya, in which two viral-vector injections carry an antigen from the liver stage of the parasite. But this might not be good enough – if a single infected liver cell goes undetected, a battalion of merozoites can burst out to target red blood cells.

BREAKING THE CYCLE

Plasmodium falciparum, the parasite responsible for the most severe form of malaria in people, spends a considerable part of its life cycle inside the human body. As the parasite takes on different forms and infects different parts of the body, it presents researchers with several distinct targets for vaccines (white boxes).

Parasite life cycle

1. A mosquito carrying *P. falciparum* injects a form of the parasite called a sporozoite into the bloodstream.

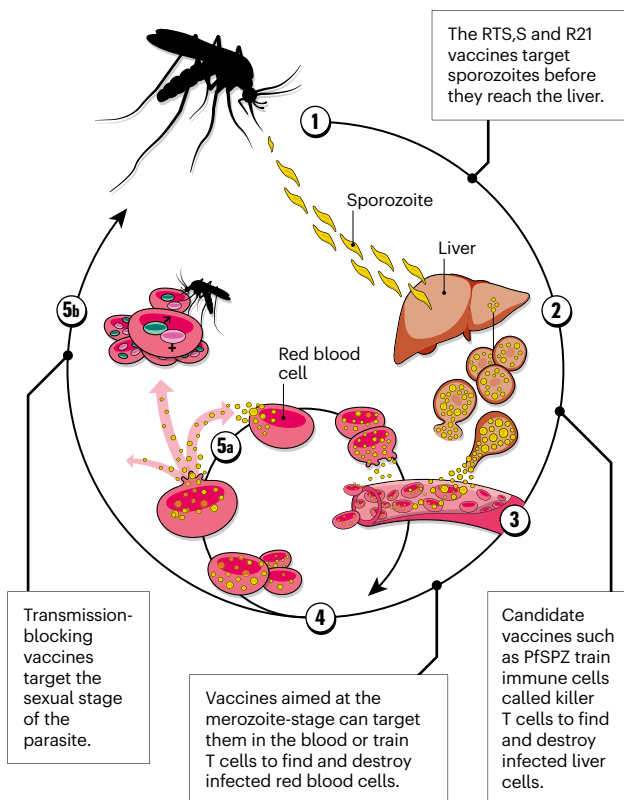
2. Sporozoites infiltrate liver cells and multiply.

3. Tens of thousands of merozoite-stage parasites burst from liver cells and enter the bloodstream.

4. Merozoites hijack red blood cells and multiply.

5a. Merozoites burst out and infect more red blood cells, releasing toxic substances that cause many of the clinical symptoms of malaria.

5b. Some merozoites mature into a sexual form. These are taken up by mosquitoes, and sexual reproduction occurs and new sporozoites are generated.



Stephen Hoffman, a malaria researcher and chief executive of biotechnology company Sanaria in Rockville, Maryland, has an idea about how to elicit a T-cell response without relying on segments of protein. In the early 1970s, research suggested that people gained immunity to *P. falciparum* if they were bitten by irradiated mosquitoes carrying sporozoites. Hoffman founded Sanaria in the 1990s to continue this work and develop a vaccine containing live, but radiation-weakened, parasites. Sanaria injected volunteers with weakened *P. falciparum* parasites that travelled to and developed inside liver cells for a few days, but then stopped replicating and did not leave the cells to cause disease. Because the parasite remained in the liver stage, immune cells learnt how to identify it in future infections.

Sanaria says that its vaccine, PfSPZ, has shown 100% protection in studies^{4,5} in which adults who have not been previously exposed to malaria, mostly in the United States, are infected with *P. falciparum*. However, in endemic regions, people can be infected by other species, such as *Plasmodium vivax*. In February, Hoffman says, a trial will begin on the island of Bioko, 32 kilometres off the coast of West Africa, where malaria is endemic. The trial will involve 2,100 people between the ages of 2 and 50, and gather safety and efficacy data for regulatory approval. “I’m pretty confident that the vaccine also works,” says Duffy. Duffy is involved in another PfSPZ study, to be based in Mali, which is now recruiting 900 women to assess the protection the vaccine offers during pregnancy. Hoffman is also planning a trial in Indonesia to see whether PfSPZ protects against *P. vivax*, and is trying to produce *P. vivax* sporozoites for testing in controlled human studies.

“Whole-parasite vaccines have shown significant promise, and we can make them even better,” says Kappe. Rather than training the immune system to recognize and respond to one parasite protein, the live parasite teaches it about lots of antigens at once, and should, therefore, generate a considerable immune response. Kappe is also working on a live attenuated malaria vaccine. Instead of hobbling the parasite with radiation, Kappe uses gene editing. A crucial DNA-packaging protein is sabotaged using the gene-editing tool CRISPR–Cas9, creating a parasite that can enter liver cells but never leave them. “Our new parasite can infect the liver, replicate and become an antigen-making machine,” says Kappe.

Kappe has shown that this strategy prevents parasites from coming out of the liver and that it triggers a potent immune response, but, so far, only for mice hosting human liver and

blood cells. He and Hoffman might collaborate to insert *P. vivax* genes into attenuated *P. falciparum* sporozoites, with the aim of providing protection against both. Together, this pair of parasites accounts for 95% of malaria.

Despite their promise, live attenuated vaccines have engendered some scepticism in the malaria-vaccine community. Some suggest that their efficacy might not be as impressive in the real world as it seems in studies that deliberately expose volunteers to malaria. In these, a defined parasite strain is introduced, says Doolan, “whereas in the field you typically have hundreds of thousands of different strains of malaria and multiple species”. Doolan also warns that gene-edited vaccines could “revert to virulence”, shaking off the shackles of attenuation and causing malaria rather than preventing it.

Whether Sanaria can make its vaccine affordable is also a subject of debate. The manufacturing process involves infecting sterile mosquitoes with the parasite, incubating them for several weeks, and then removing their salivary glands by hand. The parasites in these tiny glands are then purified and packaged. “It’s obviously a tedious process, and will make the vaccine more expensive,” says Good.

“Many people say it is crazy to make a malaria vaccine in mosquitoes, but we have made vaccines in pus and poo.”

But improvements in production methodology are possible. “If the vaccine works, there will be research to speed up the process,” Good says. Hill takes the opposing view, saying that it is “non-manufacturable to even a few thousand people without huge expense”. Storage and administration are problematic: in its current form, the vaccine must be transported in dry ice and delivered intravenously.

Hoffman brushes off cost criticisms, saying that vaccines such as Prevnar, used to protect against pneumococcal bacteria, take longer to make and require more arduous quality control. He also says that it might be possible to reduce the number of parasites required for a single vaccination. Each dose of standard PfSPZ harbours 900,000 sporozoites. But when volunteers also took antimalarial medication, a vaccine containing only 50,000 sporozoites gave superior results – all of the nine participants exposed to the parasite were protected⁵.

Kappe is also confident about the prospects of live attenuated vaccines. “Many people say it is crazy to make a malaria vaccine

in mosquitoes,” he says, “but we have made vaccines in pus and poo.”

During the life cycle of malaria parasites, the microorganisms exit the liver and hijack red blood cells. Once inside, merozoites multiply inside the cells and burst out. That can lead to seizures, severe anaemia and coma. Researchers have long targeted this part of the malarial life cycle with vaccines, but have had little success.

In the 1960s, hopes were raised when it was shown⁶ that parasite load could be dampened by giving someone blood serum from an adult living in a malaria-endemic region. But trials focusing on the blood phase of the life cycle fell flat – researchers targeted vaccines at a small number of surface antigens within which merozoites have evolved extreme variability to outwit their host’s immune system. “The targets of some of the earlier vaccines had a lot of genetic diversity, meaning they looked different in every parasite,” says Simon Draper, an immunologist at the University of Oxford, UK.

To avoid this problem, Draper is looking at a protein that *P. falciparum* uses to harpoon red blood cells, called RH5. “We know RH5 is highly conserved. It looks nearly identical in every parasite,” says Draper. His group currently targets the entire protein, but it is redesigning the molecule to elicit a stronger immune response and putting RH5 on a virus-like particle to ramp up the immune response. A trial to study safety and efficacy in UK volunteers showed highly promising RH5 antibody responses. A trial in Tanzania, where malaria is endemic, finished in July, and a second is on the cards for 2020.

In Australia, Good led the first clinical evaluation⁷ of a whole-parasite blood-stage vaccine in 2018. Merozoites were grown in red blood cells, disabled with an antimalarial drug, and then injected into eight volunteers. “The volunteers all developed a good T-cell response,” says Good. His group is now infecting vaccinated volunteers with malaria to see whether this approach limits the number of parasites in their blood. Again, Hill sees manufacturing problems ahead – growing parasites in blood cells will be just as difficult as carving them out of mosquitoes’ salivary glands, he says.

Stopping the spread

Vaccinating against the malaria parasite is “an incredibly complex problem, like sending someone to the Moon”, says Kappe. He thinks that greater investment is required, but this is not likely to be forthcoming. Parasite-targeting vaccines commonly fail to excite pharmaceutical companies – few have invested substantially in malaria-vaccine research. And there is a feeling among the researchers *Nature* spoke to that philanthropic funding organizations have begun to place

more attention on tackling the mosquitoes that carry the disease and preventing transmission, rather than priming people to fight off the parasite. Hill thinks eradication of malaria is not possible with the tools currently available, and that new vaccines are ultimately necessary. Duffy agrees: “An effort like elimination means you have to bring lots of tools together,” he says. “A vaccine can be an important addition.”

Vaccines could be useful in the battle to prevent transmission, too. One way to stop the parasite jumping between human and mosquito is to sabotage the parasite’s sexual stage. Duffy’s team has developed a transmission-blocking vaccine that targets the sexual forms that are taken up by mosquitoes when they ingest human blood cells. In an initial field trial of this type of vaccine, conducted in Mali in 2018, it proved safe⁸. Next, Duffy is planning to report the results of a trial of a vaccine that combines Pfs230, a protein on the outside of the parasite gamete, with GlaxoSmithKline’s adjuvant AS01, which is included in the RTS,S vaccine to boost T-cell response. A major challenge with these transmission-blocking vaccines, however, will be developing the methods to measure any reduction in malaria transmission in the field, and to prove that the vaccines work.

Hill doubts that a stand-alone transmission-blocking vaccine will ever emerge, but can see it being part of a multi-component vaccine. A combination vaccine that includes a transmission blocker could be more effective and practical.

The need for a malaria vaccine is not diminishing, and researchers are optimistic. “We can see the top of the mountain now,” says Kappe of his team’s work. Although scientists might not agree on the best approach to take, they can see progress being made – not least the wide distribution of the RTS,S vaccine. “Our best products will be combined in different ways,” says Duffy. “RTS,S has found a very specific role to reduce clinical malaria in children.” It might not be perfect, Hill says, but malaria is such a horrendous problem that even a partially effective vaccine could make a big difference.

Anthony King is a freelance science writer based in Dublin.

1. RTS,S Clinical Trials Partnership *Lancet* **386**, 31–45 (2015).
2. Venkatraman, N. et al. Preprint at medRxiv <https://doi.org/10.1101/19009282> (2019).
3. Ogwang, C. et al. *Sci. Transl. Med.* **7**, 286re5 (2015).
4. Seder, R. A. *Science* **341**, 1359–1365 (2013).
5. Mordmüller, B. *Nature* **542**, 445–449 (2017).
6. Cohen, S., McGregor, I. A. & Carrington, S. *Nature* **192**, 733–737 (1961).
7. Stanisic, D. I. et al. *BMC Med.* **16**, 184 (2018).
8. Sagara, I. et al. *Lancet Infect. Dis.* **18**, 969–982 (2018).

Jeffrey Bethony: Taking on worms

About two billion of the world’s poorest people are infected with parasitic worms. Treatments are available, but Jeffrey Bethony, a microbiologist at George Washington University in Washington DC, explains why only vaccines can eradicate infection.

Why is it more difficult to develop a vaccine for parasites than for many viruses?

Parasites go through a series of life stages and occupy several different niches in the body. They’ve also developed clever mechanisms to evade the immune system. So parasitic infections are the ultimate challenge.

Which diseases caused by parasitic worms do we most need a vaccine for?

Schistosomiasis results in the greatest disease burden, especially in sub-Saharan Africa and Brazil. There is an effective treatment, praziquantel, but without a vaccine we can’t prevent reinfection, so it has proved impossible to eliminate the parasite in low-income countries. And more than 500 million people are infected with hookworms, which can impair physical and mental development.

What progress are you making on a schistosomiasis vaccine?

We are developing a candidate vaccine based on proteins found on the outer surface of the worms. We’re currently running a phase II trial in Uganda, funded by the US Department of Defense, that targets a fragment of such a protein on the worm *Schistosoma mansoni*. A group at Leiden University Medical Centre in the Netherlands is working on a controlled human infection model, or CHIM, for schistosomiasis. This approach allows researchers to give people an experimental vaccine and then challenge them with a dose of pathogen. This would allow us to use fewer volunteers and reduce the costs of trials.

What vaccine strategies are you developing against hookworms?

We have developed two subunit vaccines, each containing a protein given with an immune stimulant. One protein degrades haem, a component of the blood protein haemoglobin.



Haem is potentially toxic, and antibodies against the degradation protein reduce the worm’s ability to eliminate haem from its blood meals. The other protein prevents hookworms from breaking down haemoglobin for consumption. We have done separate phase I trials of our two vaccines in the United States, Brazil and Gabon, and have just received funding to test both proteins using a controlled human infection model in endemic areas of Brazil. We then plan to test the simultaneous delivery of both proteins in a single vaccine.

So a CHIM study involves injecting healthy people with parasites?

Yes. We borrowed the idea from malaria researchers. I immunize people against hookworms by administering the proteins in the subunit vaccines, and then challenge the volunteers with hookworms. If the vaccines don’t work, we can get rid of the infection with drugs. If we had to wait for people to get infected, studies would take longer and cost more; CHIM studies accelerate vaccine development.

Isn’t it tricky to get people to volunteer to be infected with a parasitic worm?

We have no problem getting volunteers. There are lots of people who think that hookworms, because they can modulate the immune system, can be therapeutic for coeliac disease, Crohn’s disease or irritable bowel syndrome. That strategy is being trialled by other researchers now. We suspect that’s why some people volunteer.

Do you have any problems getting funding?

People who need a vaccine against parasitic worms can’t afford to pay hundreds of dollars for it. So there’s not lots of money spilling around. Malaria is better funded than disease caused by parasitic worms – it’s considered more important. Malaria researchers are usually one or two steps ahead of us. But success with a malaria vaccine would help all of us. If they can do it, so can we.

By Anthony King

This interview has been edited for length and clarity.

more attention on tackling the mosquitoes that carry the disease and preventing transmission, rather than priming people to fight off the parasite. Hill thinks eradication of malaria is not possible with the tools currently available, and that new vaccines are ultimately necessary. Duffy agrees: “An effort like elimination means you have to bring lots of tools together,” he says. “A vaccine can be an important addition.”

Vaccines could be useful in the battle to prevent transmission, too. One way to stop the parasite jumping between human and mosquito is to sabotage the parasite’s sexual stage. Duffy’s team has developed a transmission-blocking vaccine that targets the sexual forms that are taken up by mosquitoes when they ingest human blood cells. In an initial field trial of this type of vaccine, conducted in Mali in 2018, it proved safe⁸. Next, Duffy is planning to report the results of a trial of a vaccine that combines Pfs230, a protein on the outside of the parasite gamete, with GlaxoSmithKline’s adjuvant AS01, which is included in the RTS,S vaccine to boost T-cell response. A major challenge with these transmission-blocking vaccines, however, will be developing the methods to measure any reduction in malaria transmission in the field, and to prove that the vaccines work.

Hill doubts that a stand-alone transmission-blocking vaccine will ever emerge, but can see it being part of a multi-component vaccine. A combination vaccine that includes a transmission blocker could be more effective and practical.

The need for a malaria vaccine is not diminishing, and researchers are optimistic. “We can see the top of the mountain now,” says Kappe of his team’s work. Although scientists might not agree on the best approach to take, they can see progress being made – not least the wide distribution of the RTS,S vaccine. “Our best products will be combined in different ways,” says Duffy. “RTS,S has found a very specific role to reduce clinical malaria in children.” It might not be perfect, Hill says, but malaria is such a horrendous problem that even a partially effective vaccine could make a big difference.

Anthony King is a freelance science writer based in Dublin.

1. RTS,S Clinical Trials Partnership *Lancet* **386**, 31–45 (2015).
2. Venkatraman, N. et al. Preprint at medRxiv <https://doi.org/10.1101/19009282> (2019).
3. Ogwang, C. et al. *Sci. Transl. Med.* **7**, 286re5 (2015).
4. Seder, R. A. *Science* **341**, 1359–1365 (2013).
5. Mordmüller, B. *Nature* **542**, 445–449 (2017).
6. Cohen, S., McGregor, I. A. & Carrington, S. *Nature* **192**, 733–737 (1961).
7. Stanisic, D. I. et al. *BMC Med.* **16**, 184 (2018).
8. Sagara, I. et al. *Lancet Infect. Dis.* **18**, 969–982 (2018).

Jeffrey Bethony: Taking on worms

About two billion of the world’s poorest people are infected with parasitic worms. Treatments are available, but Jeffrey Bethony, a microbiologist at George Washington University in Washington DC, explains why only vaccines can eradicate infection.

Why is it more difficult to develop a vaccine for parasites than for many viruses?

Parasites go through a series of life stages and occupy several different niches in the body. They’ve also developed clever mechanisms to evade the immune system. So parasitic infections are the ultimate challenge.

Which diseases caused by parasitic worms do we most need a vaccine for?

Schistosomiasis results in the greatest disease burden, especially in sub-Saharan Africa and Brazil. There is an effective treatment, praziquantel, but without a vaccine we can’t prevent reinfection, so it has proved impossible to eliminate the parasite in low-income countries. And more than 500 million people are infected with hookworms, which can impair physical and mental development.

What progress are you making on a schistosomiasis vaccine?

We are developing a candidate vaccine based on proteins found on the outer surface of the worms. We’re currently running a phase II trial in Uganda, funded by the US Department of Defense, that targets a fragment of such a protein on the worm *Schistosoma mansoni*. A group at Leiden University Medical Centre in the Netherlands is working on a controlled human infection model, or CHIM, for schistosomiasis. This approach allows researchers to give people an experimental vaccine and then challenge them with a dose of pathogen. This would allow us to use fewer volunteers and reduce the costs of trials.

What vaccine strategies are you developing against hookworms?

We have developed two subunit vaccines, each containing a protein given with an immune stimulant. One protein degrades haem, a component of the blood protein haemoglobin.



Haem is potentially toxic, and antibodies against the degradation protein reduce the worm’s ability to eliminate haem from its blood meals. The other protein prevents hookworms from breaking down haemoglobin for consumption. We have done separate phase I trials of our two vaccines in the United States, Brazil and Gabon, and have just received funding to test both proteins using a controlled human infection model in endemic areas of Brazil. We then plan to test the simultaneous delivery of both proteins in a single vaccine.

So a CHIM study involves injecting healthy people with parasites?

Yes. We borrowed the idea from malaria researchers. I immunize people against hookworms by administering the proteins in the subunit vaccines, and then challenge the volunteers with hookworms. If the vaccines don’t work, we can get rid of the infection with drugs. If we had to wait for people to get infected, studies would take longer and cost more; CHIM studies accelerate vaccine development.

Isn’t it tricky to get people to volunteer to be infected with a parasitic worm?

We have no problem getting volunteers. There are lots of people who think that hookworms, because they can modulate the immune system, can be therapeutic for coeliac disease, Crohn’s disease or irritable bowel syndrome. That strategy is being trialled by other researchers now. We suspect that’s why some people volunteer.

Do you have any problems getting funding?

People who need a vaccine against parasitic worms can’t afford to pay hundreds of dollars for it. So there’s not lots of money spilling around. Malaria is better funded than disease caused by parasitic worms – it’s considered more important. Malaria researchers are usually one or two steps ahead of us. But success with a malaria vaccine would help all of us. If they can do it, so can we.

By Anthony King

This interview has been edited for length and clarity.



Bumblebees (*Bombus terrestris*) can pass on immunity to their offspring.

Generation game

Many vertebrates pass short-term immunity to their offspring, but plants and invertebrates take the process to greater extremes. **By Brian Owens**

An organism's immune response to attack is usually considered to be a personal battle. A pathogen or parasite attacks, the organism mounts a defence, and one of them wins. But sometimes, the target's relatives get involved. Many species have the ability to prepare their offspring to meet immune challenges that they themselves have faced. "Parents can somehow transfer the immunological memory of what they experienced during their lives to their offspring," says Olivia Roth, an evolutionary ecologist at the GEOMAR Helmholtz Centre for Ocean Research in Kiel, Germany.

One way in which intergenerational protection can manifest is by parents directly providing active immune components such as antibodies to their offspring. In mammals, including humans, mothers transfer

antibodies through the placenta or in their breast milk to help protect their offspring from disease early in life. Many other vertebrates, such as birds and fish, transfer antibodies to their offspring by depositing them in their eggs. And it's not just mothers – in syngnathid fishes, such as pipefish and seahorses, the males that incubate the eggs can also pass on immune components.

"It's there to protect the offspring until their immune system gets up and running," says Ben Sadd, an infectious-disease ecologist at Illinois State University in Normal. The vertebrate adaptive immune system – which can recognize and remember specific threats – needs a lot of time to mature. By providing antibodies that can spot invaders, parents can protect their offspring during the early stages of life.

The antibody-producing adaptive immune

system is unique to vertebrates. Some insects pass on antimicrobial peptides in their eggs, but invertebrates and plants have found another way to give their descendants a leg-up against the pathogens and parasites they might encounter: they boost the non-specific physical, chemical and cellular defences that make up their offspring's innate immune system.

In contrast to the short-lived protection afforded by transferring antibodies, enhancement of the innate immune system can provide lifelong protection – not only to immediate offspring but also to subsequent generations. It isn't clear how this is mediated (heritable alterations to gene expression have been proposed), but researchers are trying to find new examples, work out how they function and determine whether the mechanisms can be harnessed to improve our lives, by boosting agriculture or fighting insect-borne disease.

Seeding immunity

Because they cannot get away from attackers or infections, it is especially important for plants to put up a fight. "Plants have a sophisticated system to detect pathogens and insects that triggers a suite of defences," says Mike Roberts, a plant biologist at Lancaster University, UK. Receptors on the wall of a plant cell can recognize molecules associated with attackers. The plant can mobilize two types of defence mechanism in response to attack. They can adapt physically, for example by creating thicker cuticles or denser thickets of hairs called trichomes to reduce the amount of tissue that herbivores can eat. And they can respond chemically, by producing molecules that poison pathogens and herbivores.

As the attack goes on, it triggers a response throughout the plant. Both physical and chemical defences can spread to unaffected parts of the plant to defend against repeat infection – a process known as systemic acquired resistance. And there are dozens of examples of this resistance being passed down to offspring, enabling them to mount a faster defence to the same attacker.

In many species, the seeds of plants attacked by a pathogen or herbivore contain higher concentrations of chemical defence compounds. In tobacco plants (*Nicotiana tabacum*), exposure to tobacco mosaic virus makes a plant's progeny more resistant not only to the virus but also to some bacteria and moulds¹. Physical defences can also be passed on – in the yellow monkeyflower (*Mimulus guttatus*), the offspring of plants that have been damaged by insect herbivores have an increased density of protective trichomes².

Like plants, invertebrates lack an adaptive immune system. The innate immune system on

which they rely was long thought to provide a fast but non-specific response to pathogens, and considered unable to use experience of previous attacks to improve protection in the future. But Roth says there is growing evidence that the invertebrate innate immune system is more complicated than that. “The classification of innate versus adaptive immunity is getting confused,” she says.

Sadd, for example, found that the innate immune system of bumblebees (*Bombus terrestris*) could provide specific protection³. He showed that bees exposed to a non-lethal dose of a bacterial pathogen had an enhanced ability to survive a potentially lethal dose of the same bacteria later. He and his colleague also found that specific protection could be passed on. Antimicrobial activity in bumblebee offspring reflects their mother’s immune experience, and it comes from factors transferred through the egg⁴. The genes responsible for producing peptides to defend against particular pathogens are turned up in bees with mothers that had been challenged by the pathogens, even if the bees themselves had not encountered the threat. “Not only is the immune system primed, it is up and running,” Sadd says.

Transgenerational innate immunity is mostly seen in invertebrates and plants, but there are some tantalizing hints in vertebrates. Roth and her colleagues have reported differences in the expression of immune-related genes in the offspring and even grand-offspring of immune-challenged pipefish⁵. And glass frogs (*Hyalinobatrachium colymbiphellum*) can transfer innate immune defences such as antimicrobial skin peptides and mutualistic microbes to their embryos⁶.

Roll of the dice

Transgenerational immune priming is widespread among plants and invertebrates, but it is not universal. Diverse species show no signs of it, and publication bias against negative results means there are likely to have been many more unreported failures to find it⁵.

Even in those species that have the ability, it is not always used. In many plants, the response is proportional to the amount of stress a parent plant experiences, says Roberts. If the plant is hit by the same pathogen once or twice, that plant will be resistant. If it faces the same challenge more than four times, then its offspring will also show resistance.

This kind of threshold probably exists because building a robust immune system can take resources away from other physiological needs, such as growth and reproduction. Plants that are primed against insect attackers have a reduced yield in the absence of insect

pressure, says Georg Jander, a plant biologist at Cornell University in Ithaca, New York.

Preparing defences against one attacker can also leave organisms vulnerable to another. In plants, chemical defences involve two hormone pathways: the salicylic acid pathway, which defends against organisms such as fungi that feed on living tissue; and the jasmonic acid pathway, which defends against those that kill the plant, and against insect herbivores. “These are antagonistic pathways – you can only activate one or the other,” says Roberts. “If your offspring are primed against one, they are more susceptible to the other.”

Bumblebees are another case in point. Sadd found that the offspring of bumblebee mothers that had been exposed to a bacterial pathogen were more susceptible to a trypanosome parasite than were bees that had not been primed against the parasite⁷.

Transgenerational immunity, therefore, mainly benefits organisms that have offspring

“If the selection pressure is sudden, it provides a way to ramp up defences.”

that are likely to face similar threats to their parents. That’s true, for example, of species with lengthy parental care or those that do not disperse far. This pattern has been observed in many laboratory experiments, but there are some exceptions – scallops, for example, inherit some immunity, but disperse their eggs into the ocean⁸. More work is needed to determine whether the pattern holds up in natural systems. “Field studies are lagging behind,” says Liza Holeski, an evolutionary geneticist at Northern Arizona University in Flagstaff.

Uncovering mechanisms

The biggest question, however, is how innate immune memory is passed on, because there are no changes to the underlying genome.

The leading candidate theory is that some kind of epigenetic mechanism adds chemical modifications to the genome that turn gene expression up or down, changing the offspring’s phenotype more quickly than could be achieved by changing their genes, says Holeski. “If the selection pressure is sudden, it provides a way to ramp up defences.”

In plants, all three main epigenetic mechanisms have been implicated: DNA methylation to silence particular genes; histone modification to change how accessible DNA is for transcription; and small RNAs that intercept and degrade messenger RNA. And Roth found that the offspring of immune-challenged parent

and grandparent pipefish express 15 genes linked to epigenetic regulation differently⁵.

There are several unanswered questions, however. “We still don’t know which genes are targeted, or how they are targeted, or even whether it is specific genes or a broader genome-wide effect that just happens to have an effect on immunity,” says Roberts.

Despite the lack of a clear understanding of how transgenerational immunity works, there are already efforts to see how the phenomenon could be used in areas such as agriculture and pest control. The simplest application might be to create a vaccine that can be sprayed on plants to produce seeds that are inherently more resistant to insects, says Jander. “It would be a way to shorten breeding times, so you don’t have to painstakingly breed for that particular trait,” he says. Some evidence suggests that treating plants with jasmonic acid induces resistance to herbivores in the next generation, but there are no products on the market yet. And there are signs that treatment with the compound β -amino-butyric acid can induce pathogen resistance in plant progeny⁸.

Joachim Kurtz, an evolutionary ecologist at the University of Münster, Germany, says transgenerational immunity in insects will have clear value for pest control, and for insect farmers. “Lots of insects are vectors for disease,” he says. “This could potentially be exploited for a control strategy.” It might be possible, for example, to prime multiple generations of mosquitoes to resist the malaria parasite.

But to anyone harbouring the idea that transgenerational immunity renders vaccines for people less necessary, Roth is quick to set the record straight. Babies who are breast fed are less likely to become unwell with minor ailments such as coughs and colds than are formula-fed babies, because they are protected by their mother’s antibodies, but that only holds for a limited time. Once they start to eat solid food, their immune system has to fend for itself. “Not everything we experience can be transferred,” she says. How to defend against pathogens is something that we must learn on our own.

Brian Owens is a freelance journalist in St Stephen, New Brunswick, Canada.

1. Kathiria, P. et al. *Plant Physiol.* **153**, 1859–1870 (2010).
2. Holeski, L. M. *J. Evol. Biol.* **20**, 2092–2100 (2007).
3. Sadd, B. M. & Schmid-Hempel, P. *Curr. Biol.* **16**, 1206–1210 (2006).
4. Sadd, B. M. & Schmid-Hempel, P. *Curr. Biol.* **17**, R1046–R1047 (2007).
5. Roth, O., Beemelmanns, A., Barribeau, S. M. & Sadd, B. M. *Heredity* **121**, 225–238 (2018).
6. Walke, J. B. et al. *Biotropica* **43**, 396–400 (2011).
7. Sadd, B. M. & Schmid-Hempel, P. *Biol. Lett.* <https://doi.org/10.1098/rsbl.2009.0458> (2009).
8. Slaughter, A. et al. *Plant Physiol.* **158**, 835–843 (2012).

Heidi Larson: A need for nuance

In 2010, anthropologist Heidi Larson at the London School of Hygiene and Tropical Medicine founded the Vaccine Confidence Project to study what was driving the growing trend in hesitancy or refusal to vaccinate, and to investigate how attitudes towards vaccines spread through and between communities. Under Larson's direction, the project has monitored media coverage and social-media discussions for a decade to understand the factors that influence people's opinion on vaccines.

Why are vaccines such a contentious topic? Vaccines lend themselves to rumours and distrust because they aim to affect everybody on the planet – that's a pretty big deal. There have been anxieties and resistance to vaccines since smallpox vaccination began in the early 1800s. And some of the issues that led to the anti-vaccination leagues in the United Kingdom in the mid-nineteenth century are still relevant today – liberty and freedom of choice, the idea that vaccines are against nature and not part of 'God's plan', and concerns about safety. Today, mistrust and rumours about vaccines travel faster and further because the communication landscape is different. There are also many more vaccines to question.

How has the discussion online changed since your project started ten years ago? It has become much more polarized and more vitriolic. For many people, vaccine anxiety is connected with deep emotions, beliefs and ideologies. When one researcher (J. Kennedy *Eur. J. Public Health* 29, 512–516; 2019) compared our Vaccine Confidence Index – a measure of public confidence – with populist leanings in several western European nations, they were significantly aligned. It's important to note, however, that there is still a very large hesitant group who are not so dogmatic – they are asking reasonable questions. Using terms such as 'anti-vax' labels people as being totally against vaccines, but it's not as simple as that. There is a spectrum of views. Some people broadly reject vaccines, but others have concerns specific to one vaccine or one ingredient – or even just object to being made to vaccinate. We need more conversations with those people who are hesitant, but willing to have a discussion.



Heidi Larson says parents need honest and accurate information.

What can be done about the spread of misinformation through social media? Social-media companies are being told to manage misinformation, but they're also being told to stay out of people's private spaces. They're not in an easy position. Their strategy has been to put the more credible information at the top of searches and news feeds.

The exercise has been eye-opening. The idea that you can delete misinformation in this web of networks, conversations and ideology is illusory. People will find another forum to spread their ideas. It's much more constructive and sustainable to get people on board than to try to take away their platform.

How can people who are hesitant about vaccines be convinced they are beneficial? Having accurate, clear and honest information is fundamental, but it's not enough to change people's minds. This is also about emotions, opinions and feelings. The worst thing to do is tell people they're ignorant or stupid. Often we're talking about people's children, and we

should remember that all parents want the best for their child.

Health authorities should make it clear that they are listening and responding to the public's questions and concerns. It's not just about putting out information – we need to take a more nuanced approach. It's frustrating when the information that health authorities provide – the things that they think the public should know – doesn't address people's questions about, for instance, the safety of vaccine ingredients.

People are anxious. Parents need a space where they can have a conversation that makes them feel more confident, and helps them to make the right decision. Doctors, nurses and vaccinators are usually too busy to do that, but there are some really interesting, creative approaches emerging to engage with people.

What form might those approaches take in practice?

It could be anything from online chat forums to putting community volunteers in waiting rooms for people to talk to. I think each setting needs to get creative about finding ways to respond to people's questions. There's a real discomfort with clinical and health-care authority being challenged, but having these conversations will help us tremendously to build comfort and trust.

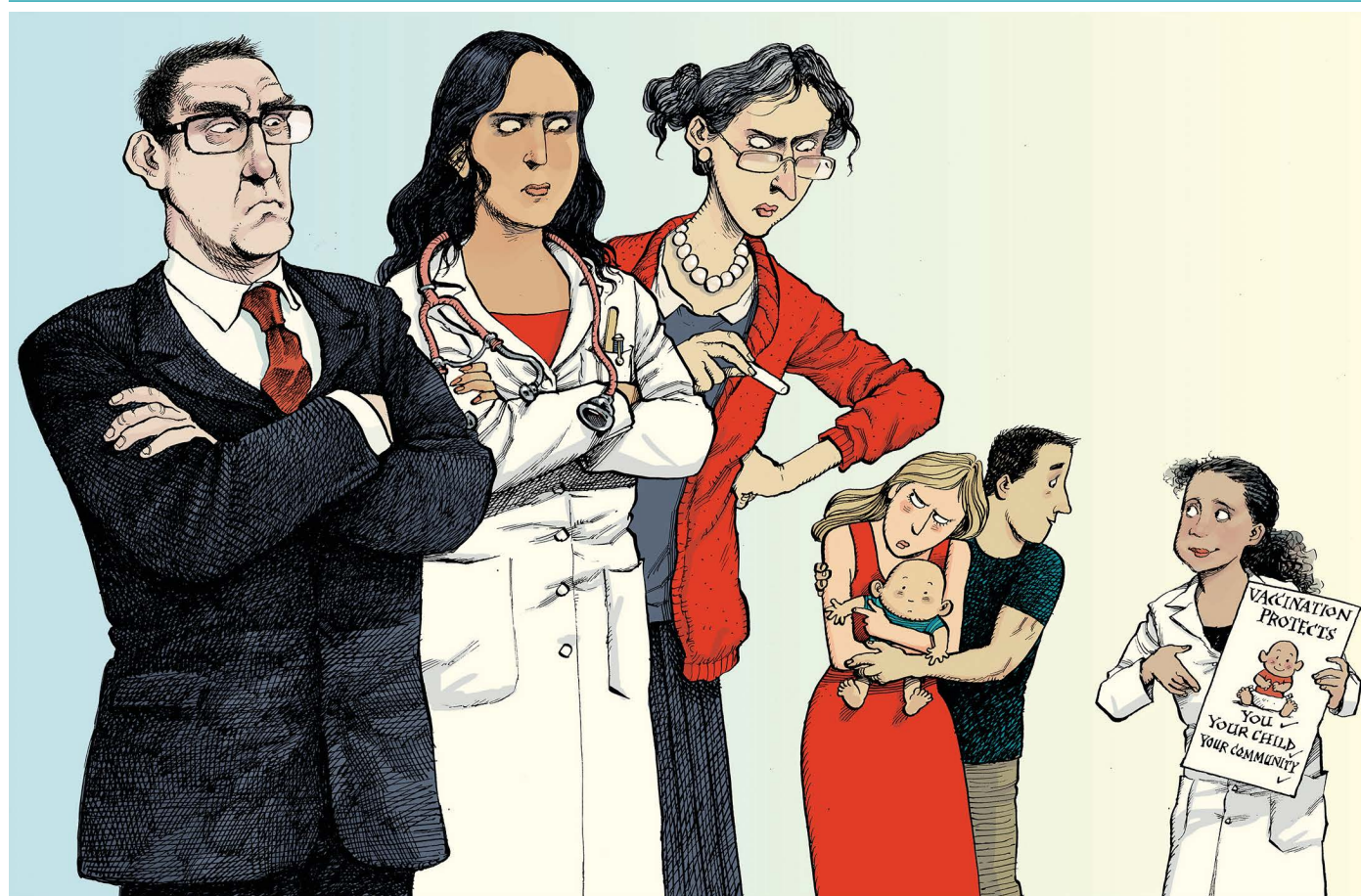
With rejection of vaccines on the rise, what does the future look like?

I hope we've seen the worst of it, but I think we might still have another hill to climb. My biggest concern is that we are going to have another very serious influenza pandemic sooner or later, and if the public opt to forgo vaccination the way they did during the 2009 swine-flu pandemic, we're in deep trouble.

That will test our ability to cooperate as a society. With levels of distrust and division as they are now, I'm not convinced we're doing too well. We've got to build a society in which people trust the system and realize that we're in it together. Or at least enough people do. We're never going to get everybody, but we can do better than we are now.

Interview by Sedeer el-Showk

This interview has been edited for length and clarity.



Forcing the issue

When people show reluctance to vaccinate their children, many countries make immunization mandatory. But not everyone favours the approach. **By Liam Drew**

In 2015, the World Health Organization (WHO) declared that the United Kingdom had eradicated the infectious viral disease rubella. The following year, it similarly designated the country as measles-free after confirmed cases numbered fewer than 125 for the second consecutive year.

Immunization rates in UK children were high at that time. They had slumped to a nadir in the mid-2000s following the false assertion in 1998 that the measles, mumps and rubella (MMR) vaccine was linked to autism. But by 2016, more than 95% of the country's 5-year-olds had received one dose of MMR, and roughly 85% had received the pre-school booster that maximizes immunity.

When 95% of a population is immune to measles, the disease cannot spread. This is known as herd immunity, and it is the cornerstone of

the WHO's long-held plan to eradicate measles globally. Achieving this would rid the world of a very serious disease, for which 1 in 1,000 cases is fatal. In 2010, eradication was considered achievable by 2020. But that time is almost here, and the disease is not close to being eradicated. In fact, it is on the rise.

During the first half of this year, Europe had 90,000 cases of measles – more than 17 times the number reported in the whole of 2016. In August, the United Kingdom lost its measles-free status (as did Albania, Greece and the Czech Republic). The United States, which is currently experiencing the highest number of measles cases since 1992, is also at risk of losing the measles-free standing that it has held since 2000.

The resurgence of measles is a symptom of falling rates of immunization against infectious

disease. “When immunization rates drop and herd immunity frays, it’s always measles that comes back first,” says Paul Offit, a paediatrician specializing in infectious disease at the Children’s Hospital of Philadelphia, Pennsylvania. “Measles is the canary in the coal mine.”

Earlier this year, the WHO named hesitancy to vaccinate as one of the ten gravest threats to global health. As a result, governments around the world are considering policies that would make vaccinations mandatory. Over the past 5 years, legislators in Australia, France and Italy have restricted school access for children who haven’t received the country’s recommended panel of vaccinations, including MMR. Some US states are doubling down on existing vaccination requirements for schoolchildren by removing the ability for parents to legally refuse vaccines for non-medical reasons. And

in September, the UK health secretary Matt Hancock responded to pressure – including a letter from four prominent London doctors calling for action to address the United Kingdom’s falling immunization rates – with the announcement that the government had taken legal advice on how it might make vaccinations compulsory.

This is a common reaction among politicians, says Noni MacDonald, a paediatrician at Dalhousie University in Halifax, Canada, and a founding member of the WHO’s Global Advisory Committee on Vaccine Safety. But mandates are not as clean a solution as policy-makers might hope. A variety of incentives and penalties have been employed, with differing levels of enforcement, and the effectiveness of each approach is not clear cut. Because the factors driving low immunization rates are not the same everywhere in the world, MacDonald says that governments should frame their policy-making decisions around two questions: “What problem are you trying to fix? And is a mandate the way to fix it?”

A pressing need

“In a better world, we wouldn’t need mandates,” says Offit. “People would educate themselves about vaccines and make the best decision for their children and for themselves. Assuming there’s not a medical contraindication, they’d get vaccinated every time.”

Evidence of vaccination’s effectiveness is resounding. Government agency Public Health England estimates that the measles vaccine, first introduced in the United Kingdom in 1968 and combined with mumps and rubella vaccines in 1988, has prevented 20 million cases of measles and saved 4,500 lives. Widely used vaccines have excellent safety records. In terms of improving public health, vaccination is second only to providing clean drinking water.

Despite this, countries around the world are failing, to varying extents, to reach levels of coverage required to achieve herd immunity – especially for MMR. Misinformation is a major problem, according to Offit. “There’s a lot of bad information out there,” he says. “It scares people – begs them to make bad decisions.”

Other researchers say that vaccines are victims of their own success. A worldwide survey published by the London-based charitable foundation Wellcome (see go.nature.com/2qg0mnp) this year showed that vaccine hesitancy is a problem mainly in high-income countries, where widespread immunization has made outbreaks of infectious disease much less common. As cases become rarer, the number of people with first-hand experience of the seriousness of the diseases diminishes. Belief in the need for vaccinations weakens, as more

people calculate that the safer course is to go without them, says Helen Bedford, a children’s health specialist at Great Ormond Street Institute of Child Health, London. “When the disease isn’t around,” she says, “half the equation has been removed – all the risk is focused on the vaccine.”

It is against this backdrop that the idea of enforcing vaccination is raised. Proponents of mandatory vaccination argue that despite what is arguably a removal of individual freedom, the ethical justification for intervention is twofold. The first argument is that the state is acting to prevent parents from making decisions on behalf of their children that unnecessarily expose them to the risk of infectious disease. Through this lens, mandating vaccination is akin to legally requiring that young children are secured in an appropriate car seat.

The second argument is that failure to vaccinate not only puts the unvaccinated individual at risk, but also anyone they come into contact with – including those too young to be immunized and people who, for medical reasons, cannot be vaccinated. “The libertarian argument falls apart,” Offit says. “If you’ve made the choice to put your child in harm’s way, and to put those who they come into contact with in harm’s way, then you’ve done harm.”

“When the disease isn’t around, half the equation has been removed – all the risk is focused on the vaccine.”

His opinion echoes that of the US Supreme Court of 1905, which upheld the legality of an 1809 mandate for smallpox vaccination in Massachusetts, stating “There are manifold restraints to which every person is necessarily subject for the common good.”

Making a mandate

Governments can never force someone to get themselves or their child vaccinated – it is a foundational principle of medical ethics that consent must be given for any procedure. The decision to make vaccination mandatory is therefore a decision to impose some form of penalty on those who do not follow the law.

A common penalty is to exclude unvaccinated children from school, because these are hotspots for disease outbreaks. This has long been the case in the United States – since 1980, all 50 states have formally linked vaccination to school entry. Australia, France and Italy have taken similar action. Australia also has legislation that withholds financial child support from the parents of unvaccinated children

without medical exemptions. In Italy, fines are also levied on parents.

But penalties can be considerably softer. Josephine Sauvage, one of the London doctors who wrote to the UK health secretary, suggests that a mandate could record children’s vaccination status at school entry, and require anyone who declines immunizations to register a conscientious objection. It would be the first such UK mandate since one was implemented for smallpox more than 100 years ago.

Although mandatory vaccination has existed in various forms for more than 200 years, there is a paucity of good epidemiological studies of the effects of different mandates, MacDonald says. The introduction of new laws is often accompanied by increased publicity about vaccination, which makes it harder to identify the specific effects of legislation. The social contexts in which mandates are applied also vary from place to place and are continually shifting.

In the United States, which recommends a panel of vaccinations, the number of states with specific mandates proliferated from 20 in 1963 to all 50 (plus the District of Columbia) in 1980. That expansion was backed by nationwide surveys in the 1970s showing that the incidence of measles was higher in states without mandates, and lowest in states where mandates were strictly enforced.

Early evidence from Italy and France shows that immunization coverage has risen with the introduction of mandates. And the No Jab, No Pay legislation withholding state benefits that was introduced in Australia in 2015 coincided with full immunization rates rising by around 3%. Nationwide coverage is now nearly 95%.

Several US states have taken steps to restrict people’s ability to opt out for non-medical reasons. In 2016, after a well-publicized outbreak of measles at Disneyland in California, the state made it impossible for people to legally opt out of immunization on anything other than medical grounds. Legislators in New York took the same action this year after a measles outbreak in Brooklyn, as did the state of Maine.

There is evidence that the California legislation has worked – between 2013 and 2017 the proportion of children attending kindergarten who were not up to date on their vaccinations halved, to 4.9%. But this might not tell the whole story. Daniel Salmon, director of the Johns Hopkins Institute for Vaccine Safety in Baltimore, Maryland, points out that the number of unvaccinated children being educated at home in California almost quadrupled between the 2016–17 and 2018–19 school years.

Salmon also contends that increases in immunization rates have been largely



A measles outbreak in April led to the New York mayor declaring a public-health emergency.

offset by a spike in the number of medical exemptions awarded since the 2016 legislation came in. There is evidence of physicians listing conditions not typically viewed as contraindications for vaccination. A further round of legislation, introduced in California in September, will see the reasons physicians give for medical exemptions monitored and controlled more closely.

The wrong problems

For Salmon, the game of legislative cat and mouse in California highlights the problems that can emerge when lawmakers try to combat a complex social phenomenon with tighter regulations. Mandates, he says, are “a quick legislative fix that will have an effect to some extent”. But to achieve stable high vaccination rates in the long term, public-health policies need to address the underlying causes of faltering uptake.

The problem highlighted by the WHO earlier this year was not vaccine refusal, but vaccine hesitancy. In most countries, the proportion of the population that staunchly opposes vaccines is less than 2%. The bigger problem, Salmon says, is the much larger group of people with some concerns about vaccination that might make them hesitant. He estimates that up to one-third of Americans have concerns about vaccines. “Making the laws stricter doesn’t address that,” he says.

The small (albeit vocal) minority of people who refuse vaccines outright rarely change their minds. The much larger hesitant population, however, does respond to information campaigns. Therefore, rather than directing a limited pot of money, health-system resources and political capital towards levying penalties for non-compliance, Salmon would prefer to

see greater investment in education and more efforts to facilitate meaningful conversations between concerned people and health-care professionals. Currently, the opportunity is limited. In the United States, Salmon says, there is no insurance code through which paediatricians can be reimbursed for consulting with parents on vaccination. And Bedford says that in the United Kingdom, the number of health visitors – the public-health practitioners who typically have such conversations – has been cut by one-third in recent years.

MacDonald agrees with the need for greater engagement. Different parents have different concerns about vaccines. For instance, some fear alleged impurities in the vaccine, whereas others are concerned about minor side effects. Studies show that public messages that broadly extol the safety of vaccines are less effective than addressing parents’ specific questions.

Bedford, however, argues that blaming falling immunization coverage on vaccine hesitancy neglects another, bigger problem: ensuring access to vaccines. This issue is commonly associated with low-income countries – and certainly, measles outbreaks last year in Yemen and Venezuela can be directly attributed to social and political events that disrupted medical services. But, says Bedford, even in high-income countries, efforts to make sure that people know how and when to get their children vaccinated are falling short. Work needs to be done, and she fears that focusing resources on implementing mandates would detract from it.

In the United Kingdom, Bedford says, vaccination rates are lowest in socially disadvantaged areas and communities in which people frequently move around. In parts of London, which has the lowest immunization

rates in the country, one in three infants change address before they’re one, meaning that the health system often loses contact with them.

For these reasons, Bedford and others argue that punitive mandates can lead to disadvantaged groups bearing the brunt of financial and social penalties. Peter McIntyre, who studies paediatric infectious disease at the University of Sydney in Australia, says that he had similar reservations when Australia hardened its stance on vaccination. Although the campaign focused on a middle-class demographic who had lodged non-medical exemptions under the old system, this wasn’t the largest group not getting vaccinated. That comprised people who were not accessing health services because of socioeconomic factors. He was concerned that denying financial support and educational opportunities to people on low incomes who were already experiencing difficulty accessing health care would only increase health disparities. Now, however, he says his fears have been at least partially allayed – the Australian government took steps to improve access by improving the vaccination register, making vaccines available to older children to catch up and investing in reminder and educational schemes.

Although less dramatic than mandates, flexible services that make appointments easier to get have increased immunization uptake. And simply sending reminders – especially for the second MMR jab, which is due around three years old when parents tend not to have as much contact with health workers – is one of the best-proven strategies for improving uptake in high-income countries. “Mandating vaccination really isn’t top of the pile in terms of what we should be doing,” says Bedford.

Indeed, most countries that achieve a stable MMR coverage of more than 95%, such as Portugal and Sweden, do not have mandates. What they have instead are populations with high confidence in vaccines, and health-care systems that provide easy access to their services.

MacDonald is wary of politicians calling for ever-fiercer laws. “They want a simple solution,” she says. “They hope that the fairy dust will fix it and they won’t have to worry any more.” But the truth is that a low rate of vaccination is too complex a problem to have such a straightforward salve. What MacDonald and many others want is careful consideration of all the factors behind low immunization rates in a community. “Everyone thinks this is a simple yes or no issue,” she says, “but it’s much more complicated than that.”

Liam Drew is a writer based in London.

ERIK PENDZICH/LAMY

Advancing vaccine innovation and public health impact at GSK



AUTHORS

Emmanuel Hanon, head of research and development at GSK Vaccines
Thomas Breuer, chief medical officer of GSK Vaccines
Rino Rappuoli, chief scientist and head of external research and development at GSK Vaccines

ADDRESS

GSK, 20, Avenue Fleming, Building W23, 1300 Wavre, Belgium

According to the World Health Organization (WHO), only access to clean drinking water rivals vaccination in its ability to help save lives (around two to three million each year). For more than 200 years, scientists have been developing vaccines to address devastating diseases and help protect people of all ages worldwide. However, public trust in vaccines is currently being eroded, and vaccine-preventable diseases have made a comeback in countries that had previously managed to eliminate them. Findings from the Wellcome Global Monitor 2018 (wellcome.ac.uk) suggest that an increased understanding of science behind vaccines tends to correlate with increased confidence in vaccines, which in turn impacts decisions about vaccination. Therefore, stakeholders in vaccination are countering vaccine hesitancy through education and by increasing the pool of easy-to-find and easy-to-comprehend, balanced information on immunization. As a leading developer of vaccines, GSK has a part to play in this, with experts who are well positioned to provide insights into the vaccines of today, and those that are to come.

SELECTING FUTURE VACCINES

Disease burden (the public health impact a disease has on a region or population) has the largest influence on which vaccines are pursued. This is often calculated in terms of quality-adjusted life-years (QALYs) and costs; considering factors such as disease frequency (is it common or rare), impact on quality of life (are the symptoms mild, severe or fatal), healthcare resource use, the financial cost to society and ability to pay. This overall burden-cost is then compared between different diseases and populations, and results can vary greatly. For instance, the developed and developing world can have differing priorities due in large part to their healthcare infrastructure, sanitation and socioeconomic circumstances.

Historically, the first vaccines were developed against diseases with high morbidity and mortality rates, such as smallpox (successfully eradicated in 1980), diphtheria and tetanus; diseases which also primarily affected children. Great societal progress was made because of advances in the fight against these diseases. In the early 1900s, families tended to be much larger (often due to the threat of high childhood

mortality) and the average life expectancy was around 60 years of age. Nowadays, families tend to be much smaller and life expectancies are around 80+ in industrialized nations. However, as societies experience a growing proportion of older adults, the burden of disease is starting to shift; of the more than 40,000 deaths from vaccine-preventable diseases that occur every year in the United States, now 99% are in adults. Therefore, thinking about vaccination across the whole life course is a key opportunity, making vaccination part of the increasing emphasis on preventative medicine for all.

THE VACCINE DEVELOPMENT PROCESS

Vaccine development is a long, costly and complex process that can often take more than 10-20 years to complete. Moreover, vaccines, unlike other medicines, are not being given to people who are already suffering from a specific illness; vaccines are given to millions of healthy individuals with the aim to help prevent them from contracting the disease. Therefore, it is critically important that vaccine developers and regulators meet the highest standards and demonstrate that the benefit of any new vaccines far outweigh any potential risks.



A scientist working on a Malaria vaccine.

Furthermore, while the very early years of vaccine development focused primarily on assessing the efficacy of vaccines, the emphasis has appropriately shifted to the benefit-risk profile of vaccines. There is the requirement that industry and regulators not only demonstrate a vaccine's efficacy, but equally its safety profile. This requires more investigations throughout the development process, even after vaccines are licensed, along with more comprehensive regulatory and licensing procedures; but all with the aim of ensuring that new vaccines have a positive benefit-risk profile – in other words that the benefits far outweigh any potential risks.

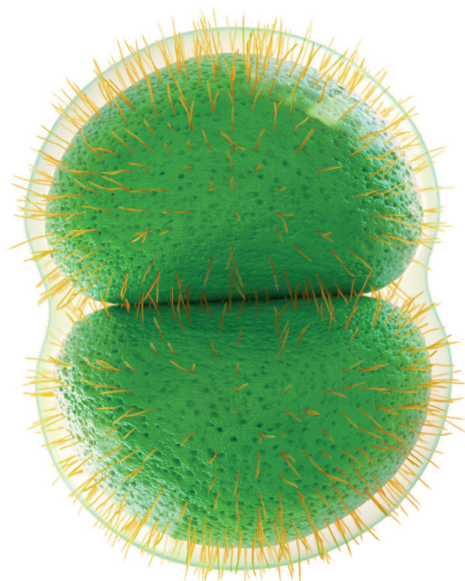


Figure 1. Meningococcal serogroup B bacteria, commonly known as meningitis B.

THE EXPLORATORY AND PRE-CLINICAL STAGES

The first step of vaccine development is the exploratory stage. This takes place in the lab, typically takes two to four years, and aims to identify natural or synthetic antigens that might help prevent or treat a disease. Historically, these antigens have included virus-like particles, weakened viruses or bacteria, weakened bacterial toxins, or other components derived from pathogens; and this stage has involved a lot of time-consuming trial and error. Today, scientific advances, and new technologies such as digitalization and artificial intelligence, are helping reduce the time needed to identify candidate vaccines and are providing solutions for vaccines that were difficult to develop in the past.

For example, in the case of meningococcal meningitis B (MenB), a rare but life-threatening illness, the exploratory stage was especially hard due to the variability of the surface proteins of MenB and the similarity between the MenB capsule and other cells in the human

body (**Fig. 1**). To overcome this roadblock, scientists under the lead of Rino Rappuoli invented reverse vaccinology, a process that screens an entire genome and uses bioinformatic tools to identify different genes as antigen candidates (Rappuoli, R. *Vaccine* **19**, 2688–2691; 2001). These are then further filtered for attributes that would potentially make safe and effective vaccine targets.

Once the antigen is identified, the preclinical stage starts. This increasingly involves using cell-culture systems (relying less and less on animal testing) to assess both the safety of the candidate vaccine as well as the immune response it causes before testing in humans is considered. The journey for many candidate vaccines ends at this stage, but those that do show promise are then approved for testing on humans.

CLINICAL TESTING

Phase I vaccine trials involve a small group of healthy volunteers (<100), although if the vaccine is for children, researchers will start with adults and gradually step down to individuals at the

target age. The aim of phase I is to test the safety of the vaccine and assess the immune response that it causes on humans. Phase II trials involve a larger group of volunteers (100–1,000), confirms formulations and doses, identifies if there is a need for boosters and determines the best intervals between each dose. This stage again evaluates safety, immune response and, sometimes, initial results on efficacy. It may also include individuals belonging to groups at higher risk of catching the disease. Phase II trials are randomized, controlled and involve a placebo group.

During phase III trials, the vaccine is tested on a much larger group of volunteers (1000s or 10000s of individuals). These tests are randomized, double blind and involve placebos (saline, a vaccine for another disease, or another substance). Phase III is an important step for identifying the less frequent possible adverse events because these might not show up in the much smaller phase I and II trials. Regulatory approval is sought before each phase of the clinical trials, with independent boards assessing the safety and efficacy of the vaccine candidate to help ensure the safety of trial participants throughout the process.

While phase III trials are running, work also begins on setting up the vaccine manufacturing facility because it can take at least six years from the time the building commences to the time that the facility typically gets its first regulatory approval. This is done to ensure a seamless start in the manufacturing and supply of a new vaccine after licensure. The construction work and clinical trials mean that the company is making large investments, at its own financial risk, before the phase III results are published and licensure is

granted by external authorities at the regional (for example by the European Medicines Agency in Europe) and/or national level (for example by the Food and Drug Administration in the US).

MANUFACTURING, TESTING AND SAFETY MONITORING

A little known but important fact about vaccines is that manufacturing a single dose of vaccine can take up to 24 months. This is mainly driven by the high number and level of quality checks that are performed on every single batch that is produced. Up to 500+ quality tests are performed on more complicated vaccines, such as multivalent pneumococcal vaccines, before release. In addition, vaccine manufacturers are regularly inspected by regulatory authorities from around the world. At GSK for example, at least one site in its global manufacturing network undergoes a routine inspection by a regulatory authority almost every week, in addition to its own internal inspection processes (GSK Annual Report 2016, p. 33; www.gsk.com).

The vaccine production process itself can be divided into the following steps, with quality checks happening throughout:

1. Generating the antigen; by growing and harvesting the pathogen's proteins, polysaccharide or deoxyribonucleic acid.
2. Releasing and isolating the antigen; by separating it from the media on which they are grown and isolating it from other proteins or growth mediums still present.
3. Purifying the antigen; by using different techniques according to protein size, physico-chemical properties, binding affinity or biological activity.

4. Adding supporting components; i.e. adjuvants (to enhance the vaccine recipient's immune response), stabilizers (to prolong shelf life) or preservatives (to allow for the safe use of multi-dose vials).
5. Filling syringes or vials; and packaging the vaccines before they are labelled and distributed worldwide.

To support post-licensure vaccine safety surveillance and pharmacovigilance, vaccine companies such as GSK employ a dedicated team who, through regular engagement activities with regulatory authorities and/or independent external experts, continually evaluate the benefit-risk profile of GSK vaccines throughout their entire lifecycle. Their work typically includes long-term effectiveness and safety studies, studies on broader populations and ongoing characterization and quantification of potential safety signals. Routine safety oversight processes, established by the authorities for medicinal products, include the independent assessment of available safety data by experts in regulatory authorities and public health institutes and aim to ensure the early detection and assessment of any potential safety signals. Both GSK and independent experts also continue to perform studies on how vaccination benefits individuals and society. This includes, but is not limited to, studies into the downstream effects of vaccines, cross-protection and broader health economic aspects.

DISEASE PREVENTION AND THERAPY

Vaccines are the product of the steady and detailed progress in our understanding of human-

pathogen interactions. Scientists today better understand how the human immune system interacts with the diseases that infect them and are using this knowledge, combined with the use of game-changing technologies, to address new disease areas (such as parasites) and populations (such as older adults), with faster response rates and shorter lead times. GSK and its partners are at the forefront of this new wave of science and technology. They are opening new possibilities for global health initiatives that aim to reduce the burden of diseases such as malaria and tuberculosis, and new ways to tackle anti-microbial resistance. These new discoveries are also starting to unleash the potential of therapeutic vaccines for diseases like chronic obstructive pulmonary disease (COPD).

COPD affects around 10% of the population >40 years of age and is the third most common cause of death worldwide, with a global death toll of 3.1 million in 2015. Prevalence of the disease is increasing and millions affected by the disease do not even know that they have it. Therefore, GSK is currently developing a vaccine targeting the bacteria implicated in 30–45% of acute exacerbations of COPD; non-typeable *Haemophilus influenzae* (NTHi) and *Moraxella catarrhalis* (Mcat). The aim of the COPD vaccine will not be to help prevent the disease itself, but rather to reduce exacerbations, slow the disease progress and hopefully improve the quality of life for the growing number of people suffering and dying from COPD worldwide.

A NEW ERA OF VACCINES

New technologies enable the industry to potentially complete tests in less time, streamline production processes, lower

costs and develop vaccine against diseases or for patient populations that were previously not possible. All while ensuring that the highest safety and quality standards are maintained. These technologies include GSK's self-amplifying mRNA (SAM), a technology that could allow our own cells to 'manufacture' the antigen in-situ, instead of introducing antigens into the body; bioconjugate technology, which could remove complexity from production of conjugate vaccines; and adjuvants.

Adjuvants are substances designed to enhance our immune response to vaccines. Although adjuvants have been used in vaccines since the 1930s, scientists today have a better understanding of how the human immune system interacts with the pathogens it confronts; including the part played by adjuvants in producing and controlling an immune response. For example, as we age, our immune system reduces its ability to mount an effective response to infection. This also means that our immune system has lower ability to mount an effective response to most vaccinations. Therefore, a new generation of adjuvanted vaccines is being designed to appropriately stimulate the immune system and thereby target diseases that primarily affect older adults, such as shingles, COPD and RSV (respiratory syncytial virus), as well as vaccines for major global health issues including malaria and tuberculosis.

PARTNERING FOR INNOVATION

In addition to technological advances, the key drivers of scientific progress are people – scientists, doctors and other experts – collaborating to find

new and innovative ways to solve problems, to work better and faster, and to maintain the highest standards of quality and safety. GSK recognizes that research and development (R&D) solutions may come from many sources and welcomes partnerships, which range from early research to late-stage development, that help to advance the frontiers of disease prevention. More than 90% of GSK's vaccines are developed in partnership and the company has more than 150 active scientific collaborations with pharma and biotech companies, consortia, charities, foundations, government researchers, academic groups and businesses in non-life sciences. GSK also provides opportunities for PhD or postdoctoral researchers to work on cross-disciplinary R&D projects for the discovery and early development of vaccines.

FROM VACCINES TO VACCINATION

The scientific progress being achieved, and solutions developed, will only benefit individuals and society if vaccines are used and turned into vaccination. In addition to developing and manufacturing cutting-edge vaccines, GSK is committed to the role it plays, together with stakeholders worldwide, in helping build understanding of and confidence in vaccination by providing clear answers to the questions being asked. For example, parents want the best for their children and may have questions about the effects of a vaccine on their child. After all, the ultimate aim of the scientific progress being made in the development of vaccines is to realize the outstanding benefits that these vaccines could potentially deliver – for our societies, and for each one of us.

nature

career guide

Cell biology



**IN
FOCUS**



MATT LINCOLN/WOLFSON BIOIMAGING/UNIV. BRISTOL

A researcher uses a confocal microscope equipped for super-resolution and fluorescence imaging at the University of Bristol, UK.

MEET THE MICROSCOPES

Advances in imaging have paved the way for new careers in cell biology. **By Nic Fleming**

Scientists are understandably wary of the word ‘revolution’. Major breakthroughs that shake the foundations of established thinking are rare compared with the small, additive steps by which science tends to progress. Yet when Werner Kühlbrandt at the Max Planck Institute of Biophysics in Frankfurt, Germany, wrote about the promise of a microscopic technique that could reveal the structure of large biomolecules at near-atomic resolution in 2014, he chose as his headline ‘The Resolution Revolution’¹.

Kühlbrandt was referring to the potential of

electron cryo-microscopy (cryo-EM), a technique that fires beams of electrons at proteins frozen in solution to reveal their structures. Cryo-EM is just one of many imaging tools driving rapid advances in cell biology and related fields. “New techniques have come on board over the last ten years that allow us to see things inside cells that we couldn’t resolve before,” says Anne Ridley, a cell biologist at the University of Bristol, UK, and president of the British Society for Cell Biology. “We can see whether two proteins are located in the same place, doing the same things, coming together

or apart in real time, understand how enzymes work and obtain structures of complexes of proteins in ways we couldn’t have dreamed of doing before.”

Scientists have been improving microscopes ever since the devices were invented in the late sixteenth century, and this process has accelerated markedly over the past decade. The 2014 Nobel Prize in Chemistry was awarded to the developers of super-resolution fluorescence microscopy, and the 2017 prize recognized the development of cryo-EM. The specifics of each microscopy technique vary hugely, as do

Meetings of minds

Symposia and conferences are good for getting updates and overviews of a field.

Researchers often have to choose between going to broad or specialized meetings. For those seeking an overview of the state of the field, the joint meeting of the American Society for Cell Biology and the European Molecular Biology Organization is by far the largest annual gathering of cell biologists in the world. Around 6,000 people are expected to attend this year's, in Washington DC on 7–11 December. Subjects to be covered will be wide-ranging, including emerging topics such as non-conventional model organisms, computational modelling and synthetic biology.

Bruce Stillman, president and chief executive of Cold Spring Harbor Laboratory in New York, will give the keynote lecture on his work on chromosome duplication in cells. There will be a variety of symposia, workshops, poster sessions and special-interest sessions. On the day before the

main meeting, there will be a full day of session on careers and professional development for academics, and a one-day mini biotech course, at which attendees can learn how scientific discoveries are turned into bioscience ventures. Other sessions will cover careers in non-profit science advocacy, science policy, outreach, scientific infrastructure management and bench-based research in industry.

There are many other options for researchers wanting to dig deeper into a particular branch of the discipline. A symposium called Seeing is Believing, for example, brings together the developers of cutting-edge imaging techniques with those applying them in the lab. This meeting attracted some 400 participants when it was last held, at the European Molecular Biology Lab in Heidelberg, Germany, in October this year. It featured sessions on the latest tools and methods transforming researchers' abilities to visualize proteins, protein complexes, organelles, cells, tissues, organs and whole organisms.



the subsets of scientists that favour them. But, fundamentally, all are ways of seeing the cell in greater detail than is possible by eye.

Each microscope technique involves a compromise. Fluorescence microscopy, in which fluorescent molecules are used to light up target proteins, cells or cellular components, allows biologists to observe live samples in real time. But because visible light cannot distinguish between objects closer than 200 nanometres to each other, it is not, on its own, enough to reveal the detailed structures of tiny functional components in cells called organelles. Electron microscopes can achieve much higher resolutions, but require a vacuum and so cannot be used on live samples. In modern science, cell biologists have access to an ever-growing armoury of microscopy tools, which can be used either alone or in combination and offer many improvements over the much older technique of crystallography (see 'Tools of the trade').

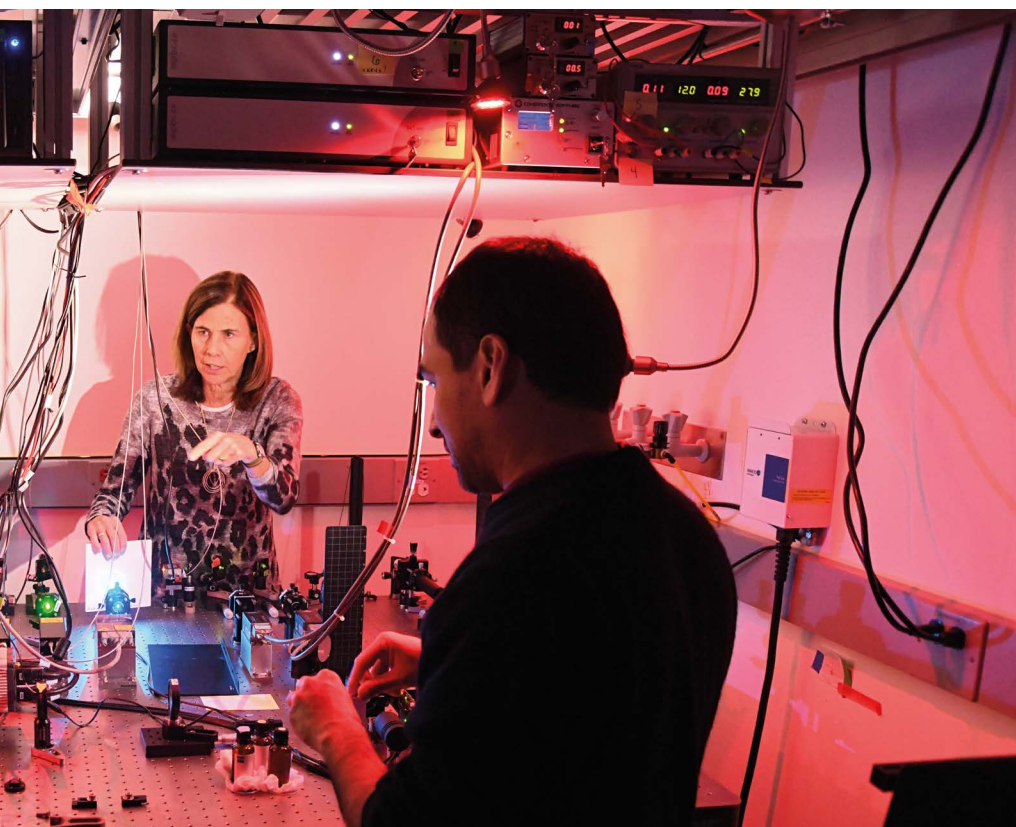
Cross-disciplinary approach

These advances have been the result of contributions from scientists in many different fields. Physicists have provided much of the technology, such as the advanced electron detectors that increased the speed and sensitivity of modern cryo-EM devices. Chemists



Delegates at the joint meeting of the American Society for Cell Biology and the European Molecular Biology Organization meeting in 2018.

PAUL SAKUMA PHOTOGRAPHY



Jennifer Lippincott-Schwartz demonstrates a microscope capable of super-resolution imaging.

have developed brighter fluorescent probes that illuminate targets for longer. Statisticians and computer scientists have improved image processing and analysis techniques. “The acceleration in imaging has come about through this incredible synergy,” says Jennifer Lippincott-Schwartz, a cell biologist at the Howard Hughes Medical Institute’s Janelia

“These techniques open up tremendous vistas for the types of questions we can answer.”

Research Campus in Ashburn, Virginia, who helped to lay the foundations for the development of super-resolution microscopy with work during the 1990s on the use of green fluorescent proteins to visualize cellular trafficking pathways in living cells².

Many advances have been made using these microscopy tools. Lippincott-Schwartz and her colleagues, for example, used a form of light-sheet fluorescence microscopy with confocal microscopy to capture 3D colour footage of the interactions between different types of organelle. “We were able to map out the relationships between six types of

organelles, how fast they were moving and the contacts they made with each other,” says Lippincott-Schwartz, whose paper³ was published in *Nature* in 2017. “That’s important if you want to understand the cross-communication between organelles, which is one of the big interests among cell biologists right now.”

The growing availability of these advanced techniques presents opportunities for early-career cell biologists. Most obviously, it increases the number of processes cell biologists can probe. “These techniques open up tremendous vistas for the types of questions we can answer,” says Lippincott-Schwartz. Structural biologist David Barford at the MRC Laboratory of Molecular Biology in Cambridge, UK, has used cryo-EM to advance understanding of some of the cellular mechanisms involved in mitosis⁴, a type of cell division that results in the formation of two daughter cells with the same chromosomes as the parent cell. “For academic scientists, the ability to determine structures at atomic resolution with electron cryo-microscopy can be very important in the design of new experiments and testing of biological hypotheses,” he says.

Barford adds that the potential benefits to early-career researchers of acquiring an in-depth understanding of the latest imaging techniques could extend beyond the

Tools of the trade

A wide variety of techniques open up multiple lines of investigation.

1

Widefield microscopes use intense light sources to illuminate entire specimens, and are less complex than other technologies.

2

Confocal microscopes use pinholes to illuminate points of interest. That cuts down on out-of-focus or background fluorescence and means they can achieve higher resolution and greater contrast than widefield microscopes.

3

Light-sheet microscopy scans samples using a very thin plane of laser light rather than a point. It has transformed developmental biology by allowing the tracking of cells and tissues in living organisms.

4

Structured illumination microscopy is a variant of super-resolution microscopy (SRM) that is useful for live cells and produces images of higher contrast than do widefield microscopes.

5

Stochastic optical reconstruction microscopy, another variant of SRM, records the positions of sets of fluorescing chemicals that are switched on and off sequentially to produce images with very high resolutions.

6

Electron cryo-microscopy can reveal the atomic structures of biomolecules such as large proteins and dynamic protein complexes. The technique’s resolutions could previously be achieved only by X-ray crystallography, which requires samples to be in crystal form.

MATT STALEY

Cell biology outside the lab



Erika Shugart had wanted to be a scientist from childhood. So she found it tough to turn her back on research in 1997, after completing a molecular biology PhD at the University of Virginia in Charlottesville. But Shugart stayed close to science, and went on to carve out a successful career in science communications and policy. She has been chief executive of the American Society for Cell Biology (ASCB) since 2016.

Why did you leave research behind?

It was a difficult choice, but once I was actually doing science, I realized that although I was fascinated by it, what I really wanted was to connect science with the public and policymakers, to work in teams and to help people understand science. I also realized that I liked projects with a defined ending, which research doesn't normally provide.

What was your path out of academia?

I put together what would now be called an out-of-academia blog. I interviewed science journalists, patent lawyers and people working in communications, technology transfer and policy, and put the results up on a website. I also learnt some web skills, which probably got me my first internship at the National Research Council of the US National Academy of Sciences, because they needed a website for a meeting.

Why have non-academic careers been such a focus at the ASCB in recent years?

A 2012 international survey, with participants mostly in the United States and mostly from the life sciences and medical sciences, showed that unemployment after a postdoc had risen from 4% in 2008 to 10% in 2012 (see go.nature.com/37eug8f). Employment in what are known as alternative careers doubled between 2010 and 2012, to 16%. Around this time, it was beginning to really hit home that most people who get a cell-biology PhD will not become a professor or stay at a university. So the society created a Committee for Postdocs and Students, whose members took matters into their own hands and started to provide help, advice and support for those moving out of academia.

How does the ASCB help those looking at careers outside academia?

We offer a wide range of talks on careers and professional development at our annual meeting (see 'Meetings of Minds'). They're generated by our committees and members, and offer people a taste of the many options out there. One of our most successful programmes is a week-long course for those considering leaving academia, with interactive sessions and speakers from biotech and pharma, and also entrepreneurs. Some 67% of attendees have gone on to jobs in industry, regulatory affairs or tech transfer. We also run careers webinars.

What trends should early-career cell biologists be aware of when thinking about their futures?

The other side of the decline in available academic positions is the uptick in jobs in industry, as well as in other sectors, such as patent law, communications and policy. There are growing opportunities to collaborate with those in other fields. We continue to see more men, white men in particular, than women and people of colour staying in academia. That's a concern for academia because diversity of backgrounds brings diversity of thought.

What careers advice do you offer early-career researchers?

I've seen a lot of CVs of people looking to make the shift out of academia. What I look for is those who go out to get extra experience and skills, whether taking a course, writing articles or volunteering at a tech-transfer office. It shows commitment to a path and initiative. There are so many enriching careers for people with advanced training in science. If people think about what it is about science that makes them passionate — maybe the writing, communicating or project management — and pursue that, they'll be able to find a satisfying career.

Interview by Nic Fleming

This interview has been edited for length and clarity.

immediate research questions they are seeking to answer. "Drug companies are becoming very keen on electron cryo-microscopy as a means to determine the structures of proteins and drug targets, so moving into it could be a very good career choice," he says. Barford also thinks these techniques will grow more important and overtake older techniques used by biologists. "It will probably supersede crystallography in the job market."

It is impossible to become proficient in the use of all or even many of the latest imaging tools. Early-career cell biologists seeking to use them need to decide whether to specialise in a particular technique, or to identify

"If you become just a user rather than a developer, it limits your future potential to contribute to the field."

collaborators who can do it for them. Ridley, who studies the role of cell migration in cancer progression, advises those doing PhDs to take up any opportunities available to them to get a flavour of the different techniques. "I would recommend that anyone doing a PhD programme with the option to do rotations in different labs and gain experience of different imaging areas to do so," she says. "Even if you don't become an expert in electron microscopy, for example, working in that area for a couple of months will give you an understanding of what it can and can't do." Barford adds that researchers who leave it to collaborators to do their imaging for them risk falling behind in other ways. "If you become just a user rather than a developer, it limits your future potential to contribute to the field through developing and advancing the technology."

One of the draws of imaging for Lippincott-Schwartz is its purity as an empirical method for acquiring knowledge. "When you are imaging, you are first observing, then generating hypotheses and then designing approaches for testing your hypotheses. It's the perfect avenue for fulfilling the scientific method." She adds that the proliferation of advanced tools has made microscopy all the more attractive as a focus for cell biologists. "It can make imaging a very creative direction to take," she says.

Nic Fleming is a science writer based in Bristol, UK.

1. Kühlbrandt, W. *Science* **343**, 1443–1444 (2014).
2. Priestly, J. F. et al. *Nature* **389**, 81–85 (1997).
3. Valm, A. M. et al. *Nature* **546**, 39–40 (2017).
4. Alfieri, C. et al. *Nature* **536**, 431–436 (2016).

Cell biology outside the lab



Erika Shugart had wanted to be a scientist from childhood. So she found it tough to turn her back on research in 1997, after completing a molecular biology PhD at the University of Virginia in Charlottesville. But Shugart stayed close to science, and went on to carve out a successful career in science communications and policy. She has been chief executive of the American Society for Cell Biology (ASCB) since 2016.

Why did you leave research behind?

It was a difficult choice, but once I was actually doing science, I realized that although I was fascinated by it, what I really wanted was to connect science with the public and policymakers, to work in teams and to help people understand science. I also realized that I liked projects with a defined ending, which research doesn't normally provide.

What was your path out of academia?

I put together what would now be called an out-of-academia blog. I interviewed science journalists, patent lawyers and people working in communications, technology transfer and policy, and put the results up on a website. I also learnt some web skills, which probably got me my first internship at the National Research Council of the US National Academy of Sciences, because they needed a website for a meeting.

Why have non-academic careers been such a focus at the ASCB in recent years?

A 2012 international survey, with participants mostly in the United States and mostly from the life sciences and medical sciences, showed that unemployment after a postdoc had risen from 4% in 2008 to 10% in 2012 (see go.nature.com/37eug8f). Employment in what are known as alternative careers doubled between 2010 and 2012, to 16%. Around this time, it was beginning to really hit home that most people who get a cell-biology PhD will not become a professor or stay at a university. So the society created a Committee for Postdocs and Students, whose members took matters into their own hands and started to provide help, advice and support for those moving out of academia.

How does the ASCB help those looking at careers outside academia?

We offer a wide range of talks on careers and professional development at our annual meeting (see 'Meetings of Minds'). They're generated by our committees and members, and offer people a taste of the many options out there. One of our most successful programmes is a week-long course for those considering leaving academia, with interactive sessions and speakers from biotech and pharma, and also entrepreneurs. Some 67% of attendees have gone on to jobs in industry, regulatory affairs or tech transfer. We also run careers webinars.

What trends should early-career cell biologists be aware of when thinking about their futures?

The other side of the decline in available academic positions is the uptick in jobs in industry, as well as in other sectors, such as patent law, communications and policy. There are growing opportunities to collaborate with those in other fields. We continue to see more men, white men in particular, than women and people of colour staying in academia. That's a concern for academia because diversity of backgrounds brings diversity of thought.

What careers advice do you offer early-career researchers?

I've seen a lot of CVs of people looking to make the shift out of academia. What I look for is those who go out to get extra experience and skills, whether taking a course, writing articles or volunteering at a tech-transfer office. It shows commitment to a path and initiative. There are so many enriching careers for people with advanced training in science. If people think about what it is about science that makes them passionate — maybe the writing, communicating or project management — and pursue that, they'll be able to find a satisfying career.

Interview by Nic Fleming

This interview has been edited for length and clarity.

immediate research questions they are seeking to answer. "Drug companies are becoming very keen on electron cryo-microscopy as a means to determine the structures of proteins and drug targets, so moving into it could be a very good career choice," he says. Barford also thinks these techniques will grow more important and overtake older techniques used by biologists. "It will probably supersede crystallography in the job market."

It is impossible to become proficient in the use of all or even many of the latest imaging tools. Early-career cell biologists seeking to use them need to decide whether to specialise in a particular technique, or to identify

"If you become just a user rather than a developer, it limits your future potential to contribute to the field."

collaborators who can do it for them. Ridley, who studies the role of cell migration in cancer progression, advises those doing PhDs to take up any opportunities available to them to get a flavour of the different techniques. "I would recommend that anyone doing a PhD programme with the option to do rotations in different labs and gain experience of different imaging areas to do so," she says. "Even if you don't become an expert in electron microscopy, for example, working in that area for a couple of months will give you an understanding of what it can and can't do." Barford adds that researchers who leave it to collaborators to do their imaging for them risk falling behind in other ways. "If you become just a user rather than a developer, it limits your future potential to contribute to the field through developing and advancing the technology."

One of the draws of imaging for Lippincott-Schwartz is its purity as an empirical method for acquiring knowledge. "When you are imaging, you are first observing, then generating hypotheses and then designing approaches for testing your hypotheses. It's the perfect avenue for fulfilling the scientific method." She adds that the proliferation of advanced tools has made microscopy all the more attractive as a focus for cell biologists. "It can make imaging a very creative direction to take," she says.

Nic Fleming is a science writer based in Bristol, UK.

1. Kühlbrandt, W. *Science* **343**, 1443–1444 (2014).
2. Priestly, J. F. et al. *Nature* **389**, 81–85 (1997).
3. Valm, A. M. et al. *Nature* **546**, 39–40 (2017).
4. Alfieri, C. et al. *Nature* **536**, 431–436 (2016).